A Reverse Vaccinology Approach to Identifying Vaccine Candidate Antigens for Bovine Trichomoniasis

A thesis submitted to the University of Liverpool for the degree of

DOCTOR OF PHILOSOPHY

by

Eleanor M. Senior

December 2020

Acknowledgements

I feel I have far too many people to thank that can be realistically put in just one section.

First I would like to thank my supervisors Dr Andrew Jackson, Professor Diana Williams and Professor Robert Hirt for their guidance throughout this process.

I would like to thank all the members of IC2 who have helped me along the way, in particular Ross, Craig and Cheng for putting up with all my questions about coding and error messages. Also Bethan and Hayley, my BBSRC buddies, whose coffee and gin trips have provided me with much needed breaks and comfort. Catherine and Stu have also been essential and have provided much technical guidance.

I would also like to thank all my friends, both from Birmingham and Liverpool who have helped me along the way, particularly Deveena for taking me on trips to Maccooleys and Ruby for all of the Skype pep talks. Also to my friends from back home, and my fellow Demon Hunters, thank you for always being there.

To my parents (and the late Chalky Senior) , I am forever in your debt for your support throughout, not only this PhD, but my previous 4 years at University and during my time at school, you've truly seen me at my best and worst, especially through my numerous operations and stressful times.

To Hugh, and indeed all the Wignalls, thank you for letting me play with the cats when I was stressed. Thank you Hugh for putting up with me, taking me to football games and letting me play Assassin's Creed. I definitely owe you at least 20 PS4 games for that!

I'm sorry I couldn't mention everyone as there really are too many people to thank for helping me get this far and I really appreciate all of your support.

Eleanor Senior

Abstract

Reverse Vaccinology of Bovine Trichomoniasis

Tritrichomonas foetus is an anaerobic flagellated protist and the causative agent of the venereal disease, bovine trichomoniasis. It is responsible for significant economic losses to farmers in several countries where the disease is endemic, and currently there is no vaccine available that can prevent infection.

The aim of this thesis was to identify putative vaccine candidates from the *T. foetus* genome using a reverse vaccinology approach. This is achieved by assembling and annotating a parasite genome sequence and evaluating putative cell-surface antigen genes for their immunogenic potential.

In Chapter 2 the *T. foetus* genome is assembled using various methods and annotated using multiple computational tools. Multiple transcriptomes are produced to identify protein-coding gene sequences. After manual curation of the genome sequence there is a repertoire of 84,706 protein –coding genes.

In Chapter 3, multiple proteomics approaches, including biotin labelling, are used in order to identify cell surface proteins and to identify differences between the two *T. foetus* life stages, trophozoites and pseudocysts. A peptide array is also used to examine the immunogenicity of selected proteins when exposed to sera from naturally infected and experimentally infected cattle. Immunogenic epitopes are identified in many of the proteins and those with the greatest number of epitopes are selected for further study.

In Chapter 4, multiple strains of *T. foetus*, along with *T. mobilensis* are grown in co-culture with an MDCK monolayer in order to identify genes preferentially expressed in the presence of host cells. Cytotoxicity and cell binding assays are also performed, which shows clear differences between the different *T. foetus* strains. These genes are crosschecked with the peptide array and original transcriptomic results to produce a shortlist of potential vaccine candidates.

In Chapter 5, the structural invariance of the gene shortlist is examined. Single Nucleotide Polymorphisms are identified between all the *T. foetus* and *T. mobilensis* strains and the reference *T. foetus* genome, with some genes appearing to be invariant. These were again cross referenced to produce a list of eight potential vaccine candidates.

In conclusion, this thesis identifies eight candidate antigens, each predicted to be localised to the cell surface. These antigens are consistently immunogenic, they are preferentially expressed in the

presence of a host, and they are not genetically polymorphic across parasite strains. This work provides a basis for the development of recombinant *T. foetus* vaccine and pipeline permanent solution to endemic trichomoniasis in livestock herds.

Contents

List of Figures

List of Tables

List of Abbreviations

AA - Amino Acid

AI - Artificial Insemination

AP - Adhesion protein

BAM - Binary Alignment Map

BLAST - Basic Local Search Alignment Tool

BP - Base Pairs

BQSR-Base Quality Score Recalibration

BSA - Bovine Serum Albumin

BSPA - Bacteroides surface protein A

BUSCO- Benchmarking Universal Single-Copy Orthologs

BVECs - Bovine Vaginal Epithelial Cells

CGR - Centre for Genomic Research

CHO- Chinese Hamster Ovary

CLP - Cadherin-like Protein

DAPI - 4‘,6-diamidino-2-phenylindole

DMEM -Dulbecco’s Modified Eagle Medium

DTT - Dithiothreitol

ECL - Enhanced Chemiluminescence

EDTA - Ethylenediaminetetraacetic Acid

ELISA - Enzyme-linked Immunosorbent Assay

FACS - Fluorescence-activated Cell Sorting

FBP - Fibronectin-binding proteins

FBS - Foetal Bovine Serum

FDA - Food and Drug Administration

FDR - False Discovery Rate

FITC - Fluorescein isothiocyanate

FMDV - Foot and Mouse Disease Virus

FPKM - Fragments Per Kilobase of transcript per Million mapped reads.

GATK - Genome Analysis Toolkit

GBP - GTP-binding proteins

GFP - Green Fluorescent Protein

GLIMMER - Gene Locator and Interpolated Markov Modeler

GMP - Guanosine 3',5'-cyclic Monophosphate

GO - Gene Ontology

GPI anchor - Glycosylphosphatidylinositol Anchor

GTP - Guanosine-5'-triphosphate

GVCF - Genomic Variant Call Format

HA - Hemagglutinin

HMM - Hidden Markov Model

HMSCs - Multipotent Human Mesenchymal Stromal Cells

HRP - Horseradish Peroxidase

IgA -Immunoglobulin A

IgG - Immunoglobulin G

IMM - Interpolated Markov Models

ISC - Iron-sulphur Biosynthesis Pathway

Kb - Kilobase

KEGG - Kyoto Encyclopedia of Genes and Genomes

KOALA - KEGG Orthology And Links Annotation

LC-MS - Liquid Chromatography–Mass Spectrometry

LFQ - Label-free Quantification

LGT - Lateral Gene Transfer

LPG - Lipophosphoglycan

NR - Non-redundant

MALDITOF- Matrix-assisted Laser Desorption/Ionization Time of Flight

MAPQ - Map Quality

Mb - Megabase

MDCK- Madin-Darby Canine Kidney

MHC- Major Histocompatibility Complex

MS - Mass Spectrometry

MZ -Metronidazole

ORF- Open Reading Frame

PacBio - Pacific Biosciences

PBS - Phosphate-buffered Saline

PCR - Polymerase Chain Reaction

PCV2 - Porcine Circovirus

PDB - Protein Database

PMP - Polymorphic Membrane Protein

PSI-BLAST- Position-Specific Interated Basic Local Alignment Search Tool

PTM - Post Translational Modification

PVDF - Polyvinylidene difluoride

QC - Quality Control

RATT - Rapid Annotation Transfer Tool

RBC- Red Blood Cell

RIPA - Radioimmunoprecipitation assay

R-P - Reverse Phase

SAM - Sequence Alignment Map

SABIA - System for Automated Bacterial Integrated Annotation

SDS - Sodium Dodecyl Sulfate

SILAC - Stable Isotope Labeling with Amino acids in Cell Culture

SNAP - Semi-HMM-based Nucleic Acid Parser

SNPs - Single Nucleotide Polymorphisms SP - Signal Peptide

STI - Sexually Transmitted Infection

UTR - Untranslated Region

TBST - Tris Buffer Saline Tween20

TMD - Transmembrane Domain

TMH - Transmembrane Helix

TMT - Tandem Mass Tagging

TVV - *T. vaginalis* Virus

V - Volts

VCF - Variant Call Format

VEC - Vaginal Epithelial Cell

VQSR - Variant Quality Score Recalibration

VSN - Variance Stabilising Normalisation

WGS - Whole Genome Sequencing

Chapter 1

Introduction

Bovine trichomoniasis is a parasitic disease of cattle and is directly transmitted during coitus. Although bovine trichomoniasis has been eradicated in the UK, it is still endemic to many areas of the world, including the US, Australia and South Africa. This parasite poses a large economic and welfare problem to farmers, particularly in countries where beef and dairy are main exports or consumed in high quantities. The eradication of this disease globally would have far reaching consequences, particularly as the economic cost of this disease is predicted to be over $500 million in the US alone [1] and is likely to be much higher due to the disease going mostly undiagnosed..

1.1 *Tritrichomonas foetus*

1.1.1 Morphology

Tritrichomonas foetus is a unicellular trichomonad that causes the venereal disease bovine trichomoniasis. It is a pear-shaped organism approximately 16.5μm long and 6.7μm wide [2] [3], although this can vary slightly by strain, where strain refers to a genetic subtype of the species. It possesses a large single nucleus, three anterior flagella [4] and one posterior attached flagellum which attaches to an undulating membrane along one side [5] (Figure 1.1). Other organelles of interest include the axostyle and costa [2] [3].

The axostyle is an intercellular projection formed by microtubule arrangements [2] [6] in a ribbon-like structure. It originates near the basal bodies, runs the length of the cell from anterior to posterior and is associated with structures such as the smooth endoplasmic reticulum and hydrogenosomes [7] (Figure 1.2) and protrudes through the posterior end. The axostyle may have a

role in the nuclear division as, during mitosis, two axostyles are present and become associated with the nuclear membrane [7]. It may also have a structural role, providing support for the flagella canal [2]. However, when *T. foetus* cells are treated with drugs that affect microtubule function, such as colchicine, the axostyle does not seem to be affected [2] [8], this may suggest that it is comprised of a different form of microtubules than the flagella and cell membrane as their morphology is altered by the addition of these drugs [9] [8]. It is thought that the axostyle is made of stable microtubules rather than labile microtubules as appears to be the case with the flagella [8]. At present, there are very few conclusive studies on the *T. foetus* axostyle and its functions.

The costa is a structure only found in trichomonads [2] and is made of striated fibres [2]. It is found in the cytoplasm [10] and runs parallel and beneath the undulating membrane [11]. It has an association with the flagella and it is thought this may be necessary for motility. It may have a role in energy production or as an energy reserve [11] as there is biochemical evidence of ATPase [10] and it is thought it is involved in cell motility.

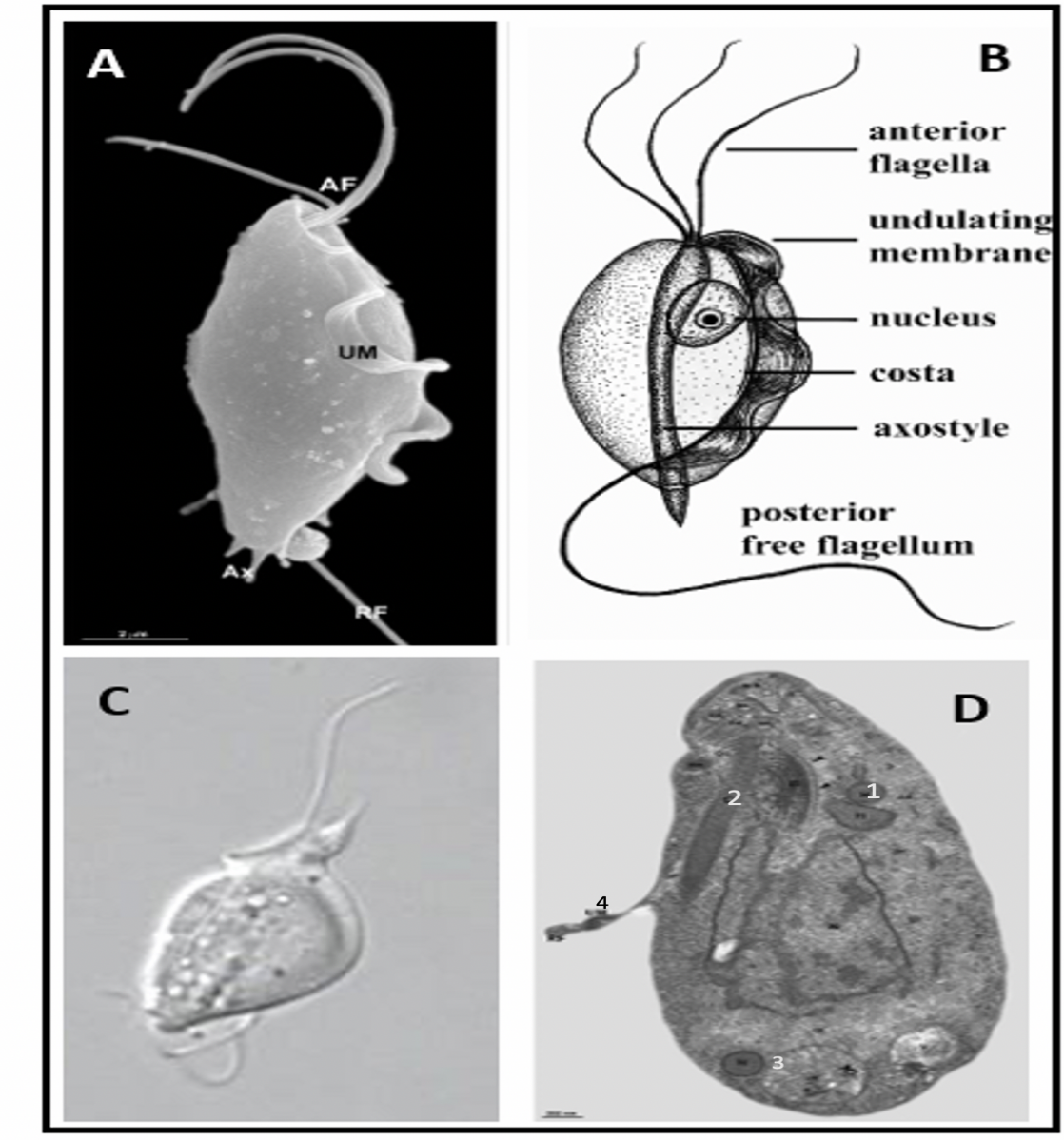

Figure 1.1: Representations of *Tritrichomonas foetus* trophozoites, showing:
(A) Scanning electron microscope image reproduced from Rosa (2013) [12] showing two anterior flagella (AF), undulating membrane (UM), axostyle (Ax) and posterior flagellum (PF
(B) *T. foetus* line drawing showing key morphological features: the flagella, axostyle, costa, nucleus and undulating membrane reproduced from K-state.edu [13] (Original line drawings by Jarrod Wood)
(C) Light microscopy of *T. foetus*, reproduced from [14] blogs.cornell.edu
(D) Transmission electron microscopy of *T. foetus* from Rosa 2013 [12] showing hydrogenosomes (1), costa (2), nucleus (3) and undulating membrane (4).

Trichomonads do not possess mitochondria [2]. Instead, *T. foetus* and other trichomonads such as

T. vaginalis have hydrogenosomes. These are hydrogen producing derivatives of mitochondria [15] that regulate the cellular metabolism [16]. In particular, they are known to function in glycogen metabolism along with the axostyle and costa [2]. Whereas mitochondria use oxidative phosphorylation to produce ATP [17], hydrogenosomes use substrate level phosphorylation and have no electron transport chain [18] but do possess an iron-sulphur biosynthesis pathway (ISC) (Figure 1.2). Reflecting its common ancestry with eukaryotic mitochondria, the hydrogenosome possesses a double membrane [3] and and can produce hydrogen and ATP from imported pyruvate produced by glycolysis in the cellular cytosol [17] [19]. They can also uptake calcium and have areas for calcium deposits on their periphery [3]. The hydrogenosomal enzymes of *T. vaginalis* are homologous with adhesin proteins [2]. The *T. vaginalis* cell surface proteome has identified enzymes that are known to localise in hydrogenosomes , as well as the cell-surface, and adhesins. The cell can secrete these hydrogenosomal enzymes and also transcription initiation factor like proteins [20], both can cause cytotoxic effects and it is likely that some similar enzymes and and secreted proteins are also produced in *T. foetus* [20].

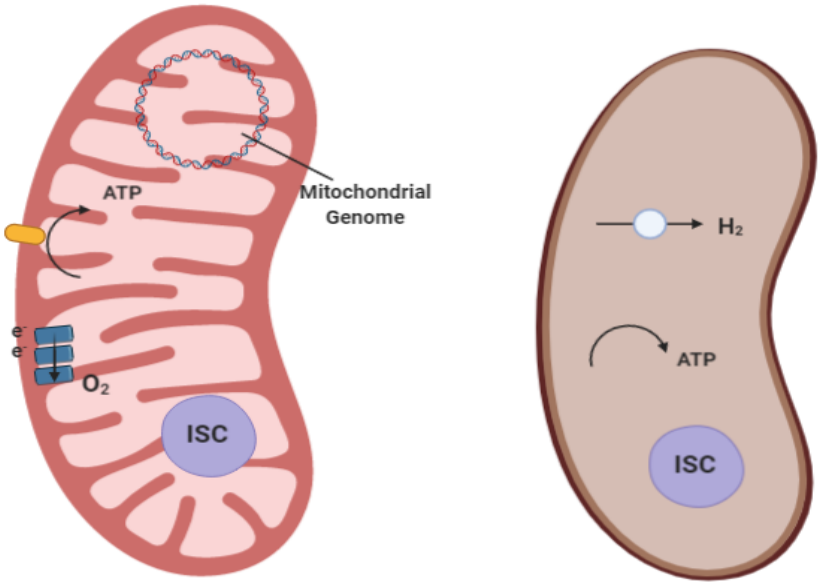

Figure 1.2: Comparison of the structures of the aerobic mitochondria and the hydrogenosome using a simplified view showing the major distinguishing features of each. The hydrogenosome does not possess a mitochondrial genome or the necessary machinery for oxidative phosphorylation. However, both the mitochondria and hydrogenosomes are double membrane bound. The hydrogenosome can use pyruvate imported from the cytosol to produce hydrogen and ATP. Adapted from Burki (2016) [18].

1.1.2 Life Cycle and Host Range

T. foetus is directly transmitted between organisms without a vector and, in cattle, is sexually transmitted from cows to bulls and vice versa during coitus leading to spontaneous abortion, foetal maceration and infertility in cows. Parsonson *et al.* (1976) [21] found that a single mating service with an infected bull infected 95% of females that had not previously borne calves. Although both bulls and cows harbour the infection, bulls are considered to be the maintenance host since once they are infected they become asymptomatic carriers for life [22]. In contrast females, when infected, are more likely to exhibit a range of clinical signs: vaginitis, endometriosis, early abortions and in some cases either transient or long-term infertility [5]. Although the infection can often be resolved in cows they can still become infected in subsequent seasons [23]. Due to the asymptomatic nature of the infection, a further difficulty is identifying and separating infected animals from the herd [24].

The age of the bull is an important factor in transmission of the parasite; older bulls (over the age of 4 years) are more likely to be infected and retain infection than younger bulls. This is thought to be due to the development of deeper crypts (microscopic invaginations in the epithelium of the penis and prepuce) in the older animals, which provide a more conducive environment for the trichomonad [25]. Susceptibility to infection has also been found to depend on cattle breed. For example, *Bos taurus* has been found to be up to 6 times more likely to retain a *T. foetus* infection than *Bos indicus* [26], and within *Bos taurus* the breeds Angus, Simmental and Charolais are more likely to have and retain the infection. It was also found that owners with a good knowledge of the disease were 70% less likely to have incidence of *T. foetus* in the herd [27]. In addition to transmission by direct contact between the bull and heifer, the disease may be spread via contaminated artificial insemination equipment [28].

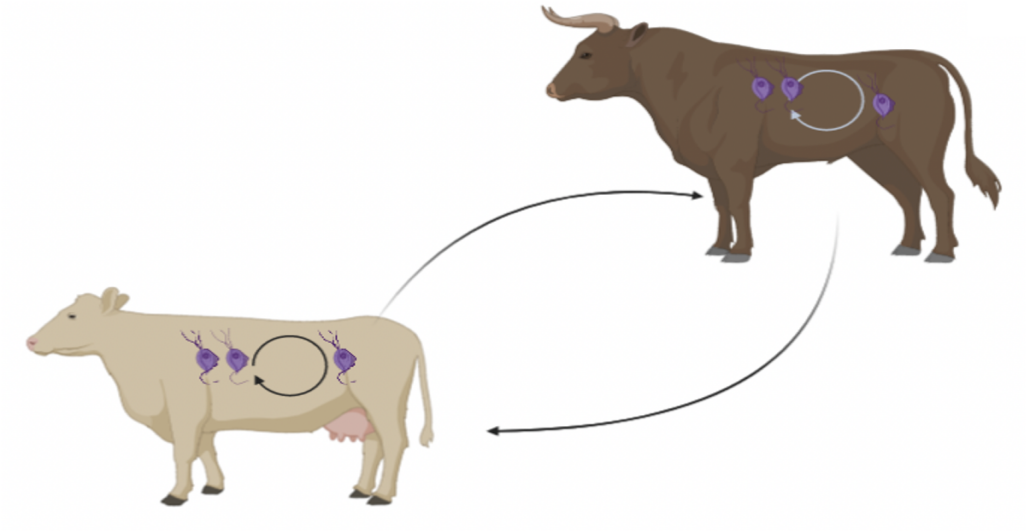

Figure 1.3: The life cycle and transmission of *T. foetus* in cattle. The parasite reproduces within the urogential tract of the cattle and is directly transmitted by coitus and has no intermediate host.

1.1.3 Life Stages

T. foetus may adopt two distinct developmental forms. Actively growing cells typically adopt a trophozoite (amoeba-like) form but can internalise its flagella and form a pseudocyst, which is more spherical in shape rather than pear shaped (Figure 1.4). This pseudocyst usually occurs under stressful environmental conditions [29]. It was previously postulated that pseudocysts are a degradative form of the parasite however, this does not now appear to be the case [16]. It is currently unknown what the true purpose of pseudocysts are, though it has been suggested that they protect against adverse envrionments, or function as an infective stage [9]. The fact that pseudocysts have been shown to survive for 48 hours in environments such as water bowls [30] provides further evidence that it has a protective function [31].

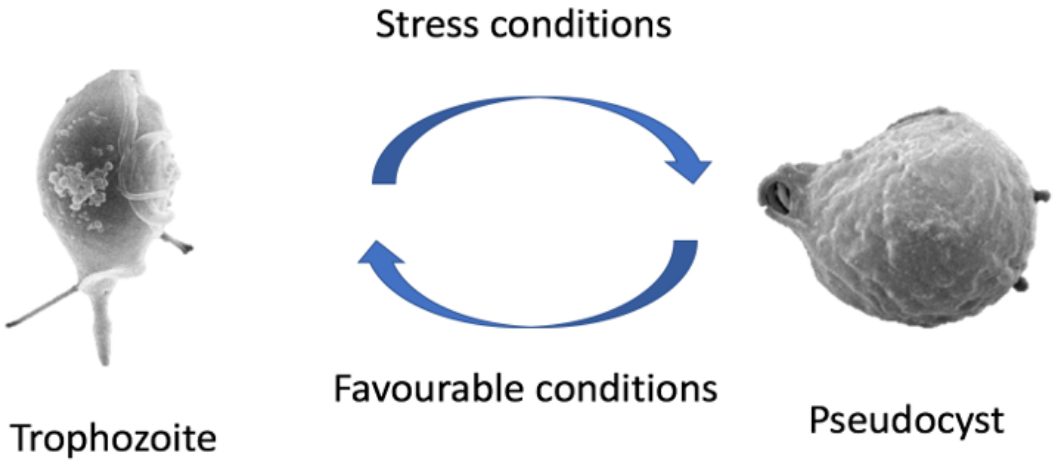

Figure 1.4: Comparison of the two *T. foetus* life stages, the pear-shaped trophozoite and the more spherical pseudocyst. Although the true reason for pseudocyst formation is unknown, it is thought that it is either an infective or protective form. The pseudocyst form can be induced by environmental or chemical stresses, such as a decrease in temperature or addition of microtubule affecting drugs such as colchicine. However, this use of colchicine is likely mimicking the morphological effect rather than inducing a true change as it is non-reversible.

1.1.4 Taxonomy

T. foetus belongs to the Phylum Sarcomastigophora, Order Trichomonadorida and the Family Trichomonadidae. Currently, the closest well-studied relative of *T. foetus* is the human pathogen *Trichomonas vaginalis* [32]. *T. vaginalis* is a sexually-transmitted infection associated with numerous health problems in humans, including: male and female infertility, vaginosis and preterm births [33]. Other related trichomonad parasites include *T. gallinae*, *Pentatrichomonas hominis* and *Dientmoeba fragilis*. Generally, trichomonads are not extensively studied and their taxonomy is not well advanced. Debate continues with respect to what is a true species, as opposed to a strain or subspecies. The taxonomic similarity between organisms, particularly when they reproduce asexually, can be used to define whether it is a species or a strain, although the deciding threshold often has to be arbitrary. New trichomonad species are often discovered, for example, in 2015 *T. gypaetinii*, previously thought to be *T. gallinae* was found to be a distinct species based on DNA sequencing [34]. Due to these new discoveries and the fact that sequencing and molecular methods are always improving, the taxonomy of trichomonads is liable to change in the future.

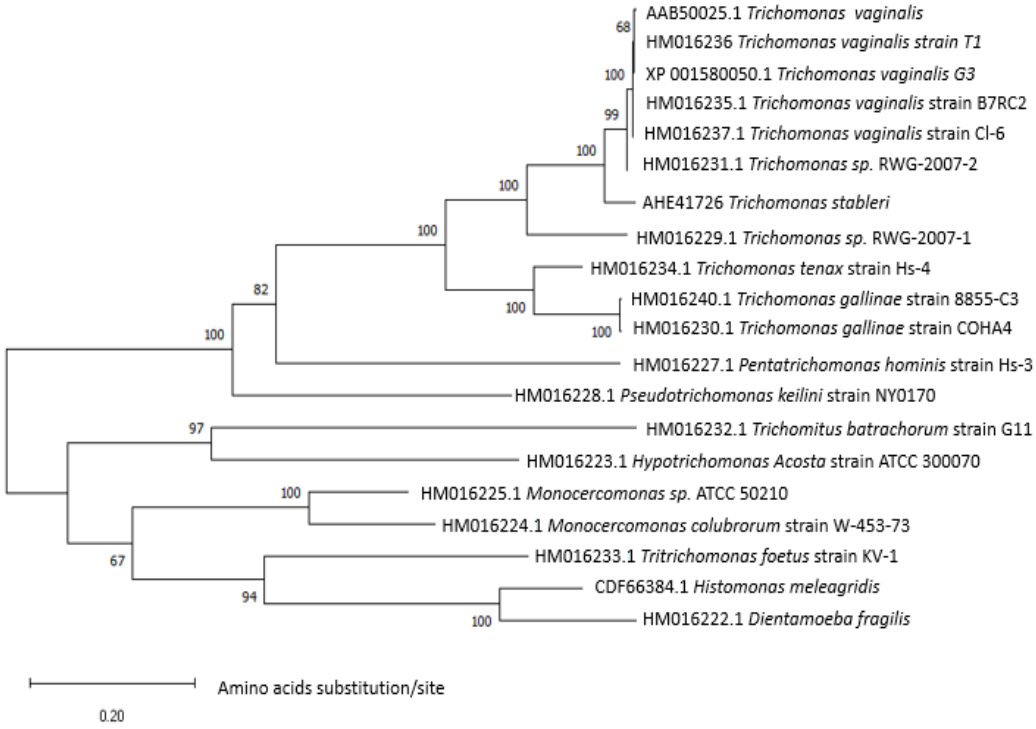

Figure 1.5: Maximum likelihood phylogenetic tree of RNA polymerase II genes from diverse trichomonads based on Maritz (2014) [32]. 20 amino acid sequences were obtained from Genbank [35] and their accession numbers are shown at the tips. The sequences were aligned using Clustal Omega. A phylogeny was estimated using Mega X [36] with a Le-Gascuel amino acid substitution model [37]. Bootstap values were inferred from 100 replicate data sets [38]. A discrete Gamma distribution was used to model evolutionary rate differences among sites. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site.

1.1.5 *Tritrichomonas foetus* in other species

Besides cattle, *T.foetus* infection is commonly found in cats and pigs and uncommonly (in three known cases), humans. Although the parasite is sexually transmitted in cattle this is not the case in other hosts. In cats, where it is an important cause of diarrhoea and colitis [39], it is transmitted by the faecal-oral route [4]. Infection is mainly found in young cats particularly in colonies and multi cat households where it is believed to be spread by contact with faecal matter.

The organism is viable in cat faeces for days and also for short periods in urine, wet cat food and water [40]. Studies have shown that 10% - 15% of cats in the UK and USA are infected. When infected, many cats are infected for life and the only treatment available is ronidazol [41]. However,

there is evidence that this drug is losing its potency due to the development of resistant strains. These drugs cannot be used in pregnant animals due to their teratogenic nature [41]. Although isolates from the feline gastrointestinal tract and the bovine reproductive tract are morphologically identical, consistent differences in their DNA have been found [4] [42]. Morin-Adeline *et al.* (2014) [42] suggest that the prominent genetic similarities between the two strains is evidence of recent adaptations to the mucosal environment of their respective hosts. However, both the feline and bovine strains appear to occupy distinct niches in the UK and there does not appear to be any evidence of sporadic cases in cattle from contaminated cat faeces. This provides some evidence that the two strains could be separate species. The name *T. blagburni* had been proposed as a possible name for the feline strain due to the differences in host specificity and pathogenicity [43].

More recently, in 2020, the rRNA genes of several American bovine *T. foetus* strains have been sequenced [44], as have suspected *T. foetus* samples from other mammalian hosts. There was over 99% similarity between the newly isolated American strains and previous bovine isolates. When they were compared, two clear clades of Tritrichomonads were seen: one containing the bovine, swine and and human isolates and one containing feline, avian, squirrel monkey and canine isolates. All Tritrichomonas species sequenced were more similar to each other than to any other Trichomonads sequenced, such as *Pentatrichomonas hominis* or *T. vaginalis*. This study concluded that all of the mammalian Tritrichomonas species that were sequenced belonged to *T. foetus*, however, they also agree that further study is needed to fully support this [44].

Tritrichmonas suis is a commensal in the intestinal tract of pigs. *T. suis* and *T. foetus* were originally considered to be separate species, but recent evidence from genetic studies, along with microscopy [3], has led to the conclusion that they are in fact the same species [45] [46]. Tachezy *et al.* (2002) [45] analysing 16S rRNA sequences , found no species specific differences and showed that they had the same cell morphology and structure. Experiments involving infecting cattle with *T. suis* and pigs with bovine *T. foetus* and seeing if there are clinical symptoms of disease could provide evidence whether the strains are the same species and can occupy the same niches. The prevalence of *T. suis* in pigs is commonly high [47]. *T. suis* was found in the nasal cavity of 55% of pigs tested and it was found in 64.5% of Australian pigs even though there were *T. foetus* negative cows nearby [48]. Additionally, there have been three isolated reports of *T. foetus* infections in immunocompromised humans [49]. There has also been a case of *T. foetus* associated with periodontitis in a 52-year-old man, again, he was immunocompromised [50].

1.2 Other Trichomonads

Trichomonads are found in essentially all species of vertebrates [51] [52] and were even thought to infect the dinosaurs [53]. Many trichomonads are simply commensals, for example, *Tritrichomonas suis* in pigs [45] and some can be ectosymbiotic, such as *Mixotricha paradoxa* [54]. However, most known species are pathogenic and can cause a range of serious diseases. There are four species that are considered to be human parasites: *Trichomonas vaginalis*, *Trichomonas tenax*, *Dientamoeba fragilis* and *Pentatrichomonas hominis*, the latter two species are also parasites of non-human animals [32], which may indicate zoonotic potential between the species. Other trichomonads, such as *T. foetus*, *T. gallinae* and *T. mobilensis* are known animal parasites and are not known to infect non-immunocompromised humans. Trichomonads are not limited to one area of the body; with *Trichomonas vaginalis* found in the urogenital tract, *Trichomonas tenax* found in the oral cavity and *Dientamoeba fragilis* and *Pentatrichomonas hominis* both found in the gastrointestinal tract [32]. As previously stated, *Tritrichomonas foetus* can be found in the urogential tract of cattle but the gastrointestinal tract of pigs and cats and *T. gallinae* is found in the oral cavity and crop of birds. There are different modes of transmission depending on the species, including faecal-oral and sexual. However, there is no known intermediate host in the life cycle of any of the trichomonads. There has been much debate on whether certain species of trichomonads, for example *T. suis* and *T. tenax* are merely variants of other trichomonads [55] [56] due to their high similarities in their rRNA and class II fumarase gene sequences [56] [57] [58]. This could mean that the taxonomy of the trichomonads could be revised in the future.

1.2.1 *Trichomonas vaginalis*

T. vaginalis is the most common non-viral human sexually transmitted pathogen [20] [59] and the causative agent of trichomonaisis. Regarding incidence: *T. vaginalis* has an incidence of 177.7 per thousand in women in the Americas and 180.6 per thousand in men. In south east Asia it is 40.3 in females and 50.1 in males [60]. It is found globally and has, in some areas, reached an prevalence of 25% or higher [61] and is associated with several other diseases and complications, such as HIV and pelvic inflammatory disease [62]. It is also seen concurrently with a range of other diseases, such as gonorrhoea and bacterial vaginosis. *T. vaginalis* has also been thought to be associated with various cancers, as have many trichomonads [63] [64] [65]. It may also be zoonotic from birds in a similar way to *T. tenax* [60]. One reason why human trichomoniasis is so prevalent is due to the fact that most men and up to 50% of woman may be asymptomatic [66] [61]. Due to this

asymptomatic nature, estimating the true prevalence of the parasite is very difficult [67]. The fact that *T. vaginalis* infects humans means that it has been studied in much greater detail compared to its bovine counterpart and the similarity between the two means that knowledge about *T. vaginalis* could be applied to *T. foetus*, such as virulence factors.

T. vaginalis has only one anterior flagellum rather than the three found in *T. foetus*. Apart from this, morphologically speaking, *T. foetus* and *T. vaginalis* are very similar [61]. *T. vaginalis* also forms stress induced pseudocysts [68] like *T. foetus* and, also like *T. foetus*, these are viable and their formation can be reversible. These these pseudocysts were found to have a reduced glycolytic metabolic pathway but the lactate dehydrogenase pathway was greatly increased. There may be similar metabolic differences between *T. foetus* trophozoites and pseudocysts. The environmental conditions in which both species interact to the host are constantly changing, for example the shedding of the epithelial lining or changes to pH of the genital tract [67] and so both would need adaptations to these environmental pressures.

Many of the relevant antigens for *T. vaginalis* that confer pathogenicity are known, such as adhesion proteins: AP120, AP65 and AP51, in addition to cysteine proteases and various lectins [67] and infection is known to produce IgG and IgA responses in vaginal fluids. Proteomics approaches have looked into *T. vaginalis* cell surface proteins along with secreted proteins that could be involved in the host-parasite interactions. In terms of host-parasite interactions; lipophosphoglycan is known to be a likely involved in the adhesion of *T. foetus*. Additionally, metabolic and membrane proteins have also been postulated as adhesin proteins.

1.2.2 *Trichomonas tenax*

Trichomonas tenax is a commensal of the human oral cavity [69] and feeds on the epithelial microbiota. It is usually found around diseased teeth and gums [70] and it is thought that infection is related to poor oral hygiene. Not a great deal is known about the organism, which older literature refers to as *T. buccalis* or *T. elongata*. It has previously been thought that it may be related to the incidence of pulmonary trichomoniasis [70], a rare infection, reported only 39 times between 1945 and 1985 [70] and thought to self-cure or was treated with antibiotics. It may be that the presence of *T. tenax* was incidental and was an opportunistic infection. The fact that most infections reported recently, were in patients with underlying conditions, such as cancers [71] further suggests that it could be an opportunistic co-infection. It is easily spread between people via saliva and droplets and the prevalence of *T. tenax* can be highly varied, ranging from 4-53% [72]. A study in

Iraq [73] showed that the incidence of the parasite was affected by many factors, such as existing oral disease, such as periodontal disease; 12.6% of patients with oral disease were positive for *T. tenax* but this dropped to 5.3% in people with no oral disease [73]. Smoking was also identified as a risk factor; the infection rate in smokers was 15.7% compared to 4.5% in non-smokers. When *T. tenax* cells were added to MDCKs (canine kidney cells), HeLa cells and 3D spheroids [69] they were found to cause cell damage and death in a similar way to *T. vaginalis* suggesting it has potential to be pathogenic or cytotoxic. It has also been suggested that *T. tenax* is a genetic variant of either *T. vaginalis* [56] or *T gallinae* [74] in a similar way that *Trichomonas suis* was found to be the conspecific with *T. foetus.* [45] [55].

1.2.3 *Dientamoeba fragilis*

Dientamoeba fragilis is another parasite of the human gastrointestinal system. It was originally classed as an amoeba but was later placed in the class Tritrichomonadea with *Histomonas* [75] [76]. When it was first discovered it was considered a harmless commensal, comparable to the views held for *Trichomonas tenax* [77]. Yet, as with *T. tenax*, there is increasing evidence that *D. fragilis* may be pathogenic [77] [78] [79], implicated perhaps in irritable bowel syndrome and diarrhoea [79]. In a study by Stark *et al.* (2010) [80], *D. fragilis* was found to be more common in the patient stool samples than *Giardia* and 83.3% of patients with *D. fragilis* experienced diarrhoea. The worldwide prevalence ranges from 0.4-32% [81] [82] and the true infection rates are likely under-reported due to a lack of accurate diagnostics and the short survival time of the parasite outside the host [81] [80] . *D. fragilis* has also been found to be related to *T. musculis* which has associations with cancer. The similarity between the organisms may mean that *D. fragilis* could possibly have the same associations with cancer [63] [83]. *D. fragilis* parasites that have similar genotypes to a human strain of *D. fragilis* have also been found in pigs, providing more evidence for the potential that is is zoonotic [84]. Little is known about parasite ransmission or life-cycle and it was long thought that there was no cyst form [84] [85]. However, Munasinghe *et al.* [86] identified a cyst form and showed that these are shed from rats, therefore providing the likely mechanism for faecal-oral transmission. Currently this is the accepted mode of transmission [75]. Stark et al. [87] showed that *D. fragilis* cysts could be found in human clinical samples, however the numbers and proportions of these were very low (0.01% of the sample) [87].

When the *D. fragilis* transcriptome was produced, it was found to be similar to that of *T. vaginalis* despite their morphological differences [76]. The trophozoites of both species were considered to

metabolically similar and *D. fragilis* was found to have a very large BspA expansion in the same way *T. vaginalis* does, suggesting the BspA family is very important amongst the trichomonads. Transcripts that were homologous to cytotoxic cysteines were also identified in *D. fragilis* [78]. These genes may provide crucial evidence in identifying the diversity between different *D. fragilis* strains [76] in a similar way to how bovine and feline isolates of *T. foetus* can be differentiated [88].

1.2.4 *Pentatrichomonas hominis*

Pentatrichomonas hominis is known to inhabit the gastrointestinal system of humans [63], in addition to many other mammalian species, including rabbits, foxes, sheep, goats and cats [89] [90] [91] [92]. As with *T. vaginalis* and some other protistan parasites may be associated with certain cancers [63] [93] [94] [64] [65], *P. hominis* may be associated with some gastrointestinal cancers [63]. In 2019, Zhang *et al.* [63] found that infection with *P. hominis* increased the risk of colorectal, oesophageal, stomach, small intestine and liver cancers. Infections with *P. hominis* increased the risk of these gastrointestinal cancers by 6.75 fold. It was previously thought that *P. hominis* was a commensal, however, as well as cancers, it has since been linked to several conditions, such as irritable bowel syndrome and rheumatoid arthritis [63] [95] [96] and may be a causative agent of diarrhoea [63]. In one case where a *T. foetus* infection was suspected in two cats presenting with diarrhoea, molecular analysis showed that the infection was due to *P. hominis* instead [92], showing that it is not likely to be merely a commensal organism. In fact, *P. hominis* may be often misidentified as *T. foetus* in clinical animal cases, such as gastrointestinal issues in cats. In a few sporadic cases, *P. hominis* has been identified in the human respiratory tract [97] suggesting it can colonise other areas. This may be similar to how *T. foetus* can occupy different niches in different mammalian hosts.

1.2.5 *Trichomonas gallinae*

Trichomonas gallinae is an emerging pathogen of birds [98]. It is found globally and has been found in a range of avian species. While it has been associated with a decline in finches in Europe, however its natural hosts are thought to be pigeons and doves (Columbidae) [98] [99]. It has been suggested that a *T. gallinae*-like parasite infected the dinosaurs [53]. The domestic pigeon (*Columba livia*) that is often credited with the wide spread of *T. gallinae* infections [100].

Prevalence of the parasite in columbiforms varies widely depending on the study, ranging from 15% [101] to 34% [102] for the same species of wood pigeon on different parts of the same peninsula. This makes determining a typical prevalence of the parasite globally very challenging. *T. gallinae* infects the crop and oesophagus of birds [103] and can cause cankers and lesions to form, leading to the death of the animal by starvation [104]. It has caused large decreases in the greenfinch population across the UK [105]. The most likely mode of transmission is direct between birds, likely via crop milk from infected parent birds to their offspring [99] [106], from raptors consuming infected animals or from courtship. Contamination of water by an infected animal has also been shown to cause spread of infection, for example, in species of chickens [99]. *T. gallinae* is also known to form pseudocysts, similar to those of *T. foetus*, which are thought to enhance infectivity or prolong longevity outside of the host [99] [107].

Several genotypes of *T. gallinae* have been described [104] [108] suggesting the possibility of multiple genotypes of the species [74]. The ITS1-5.8S-ITS2 regions of parasites isolated from racing pigeons in Poland contained two different sequence types [108]. Sansano-Maestre *et. al* [109] also investigated the ITS1-5.8S-ITS2 regions of *T. gallinae* isolates from columbiforms and raptors; they identified two genotypes, called A and B. They also showed that genotype A was more prevalent in columbiforms, whilst genotype B was more prevalent in raptors. Moreover, the genotype B was present in all isolates that were found to have lesions, suggesting that this form may have a higher virulence or pathogenicity [109]. Grabensteiner [110] discovered a further four genotypes when comparing samples from 16 birds, including pigeons, vultures and doves. The strains had high sequence identity, ranging from 84.7-97.6%. This study also demonstrated that the sequence identity between different *T. gallinae* strains and *T. vaginalis* ranged between 87.2-99%, and between *T. gallinae* and *T. tenax* of 86.6-97.3%; [110]. One strain did not appear to group closely to any of the others. Grabsteiner [110] postulated that this could be an early step in the establishment of a new strain. Due to the poor taxonomy in trichomonads, it is possible that species diversity is massively underestimated when it is based on morphological similarity. In many cases it may be that two organisms with low sequence identify are two separate species rather than separate strains.

1.2.6 *Trichomonas mobilensis*

Trichomonas mobilensis infects the gut lumen of squirrel monkeys [103]. It is known to be highly invasive and can infect up to 100% of colony members in large monkey colonies. When *T. mobilensis* cells were incubated with RK-12 cells (rabbit kidney) to investigate their virulence, they were

found to damage the monolayer. They also adhered to the flask and produced a haemagglutinin which was found to have similar properties to sialic acid-specific lectins [103] [111]. This haemagglutinin appeared to be inhibited by the presence of sialic acid. The same result occurred when the experiment was repeated using *T. foetus* parasites. Similarly, Demes *et al.* (1989) [112] added *T. mobilensis* parasites to a CHO cell monolayer, the parasites adhered within 5-15 minutes. Cell adherence was time and parasite concentration dependant and could be inhibited by sialic acid and sialyllactose. The parasites infected the luminal crypts, in a manner reminiscent of *T. foetus* infecting the crypts of the bovine prepuce. There was also evidence of mucosal invasion, this is commonly seen in *T. vaginalis* and *T.foetus* infections [103]. These results show that both *T. foetus* and *T. mobilensis* use sialic-specific lectins in some way for host-cell attachment.

When Midlej [113] compared the binding of *T. mobilensis* and *T. foetus* the binding capacity of *T. mobilensis* was lower. Feilessen [55] suggested that both *T. mobilensis* and *T. foetus* could be the same species, with *T. mobilensis* being a variant or subtype of *T. foetus*.

1.2.7 Genetics and Omics of Trichomonads

Trichomonad genetics

The number of chromosomes in the trichomonads varies with species [114], *T. vaginalis* for instance has 6 chromosomes [115]. The same karyotype was found across various different isolates, showing that it is broadly conserved. Other trichomonads, such as, *T. foetus* and *T. augusta* have 5 predicted chromosomes [114], *T. tenax* and *P. hominis* also have 6 chromosomes each. There is little known about the regulation of trichomonad genomes. In *T. vaginalis*, two thirds of genes encode transposable elements and there is evidence of a large genome expansion. Very few genes have introns. There is no evidence of TATA box elements though there is an Inr promoter element. This appears to have evolved early so likely to be a feature of all trichomonads [116]. Iron is thought to play a crucial role in gene expression as well as other key functions within the cells. This will be further discussed later in the chapter. Isoenzyme electrophoresis showed different levels of enzyme polymorphisms between trichomonad species[117]: *T. gallinae* had 5 polymorphic loci out of 11 tested whereas *T. vaginalis* had 2. Additionally, the *T. foetus* KV1 strain had malate dehydrogenase differences compared to all other *T. foetus* strains examined. Regarding trichomonad reproduction [118], it was initially assumed to be sexual due to diploidy. Xu (1998) [119] stained *T. foetus* and *T. suis* chromosomes and identified five pairs, however, segregation and recombination were found to be absent in natural populations [118]. Therefore, trichomonads appear to reproduce by

clonal reproduction. Zubacova (2008) [114] stated that all trichomonads were haploid and that *T. foetus* had five chromosomes. Even in 2020 [120], it has not been fully determined whether *T. foetus* is haploid or diploid. In *Giardia*, zymodermes sampled in many geographical areas and from different hosts, including humans and cats, were found to be almost identical [121][122] and *T. foetus* populations are similarly thought to have clonal structures.

Omics

Trichomonads have a smaller repertoire of genomic, transcriptomic and proteomic resources available compared to other protists, such as *Plasmodium* [123]. In part this is because trichomonads have large and repetitive genomes that have been difficult to reolve until recently. The advent of long-read sequencing platforms such as PacBio and Nanopore has made sequencing trichomonad genomes practical. *T. vaginalis* has a 160Mb genome and other trichomonads may be similar if they too contain a large proportion of repetitive regions. However, the range of genome sizes predicted for the parabasalids is large, with *T. tenax* having a predicted size of 133 Mb (+/- 4) and *Pentatrichomonad hominis* having a predicted genome size of 94Mb (+/-8) [114].

RNA-Seq analysis of gene expression has been used in several protists, including *T. vaginalis*. This exposed a large family of BspA proteins and also genes that were differentially expressed in the presence of iron [123]. RNA-Seq in *Pentatrichomonad hominis* identified 442 hydrogenosomal proteins [124] including multiple homologues of *T. vaginalis* hydrogenosomal proteins. Proteomics and metabolomics have also been used to study *T. vaginalis* and has been used to validate and improve *in silico* predictions [123]. *In silico* analysis of the *T. vaginalis* genome predicted 138 hydrogenosomal proteins [125] but mass spectrometry of hydrogenosomes predicted 569 proteins [126]. Additional proteomic studies could help to identify further proteins that are localised to the *T. vaginalis* organelles. In a similar way, proteomic studies in conjunction with *in silico* predictions could produce a full repertoire of *T. foetus* proteins.

1.3 The Reproductive System of the Cow

T. foetus can adapt to a range of mucosal environments. The urogenital tract of the cow is one such environment and poses several challenges to the parasite, including pH changes and an abundance of micobiotic flora and fauna. The reproductive system of a cow consists (briefly) of two ovaries, two oviducts, cervix, vagina, vulva, two uterine horns and a uterine body. Cows typically have a

21-day oestrus cycle [127], meaning that every 21 days cows are able to ovulate and conceive. For most cows their gestation period is 283 days [128] however, this can vary depending on the breed of the cow or the sex of the calf and can range from 279 to 287 days. Cows are usually able to become pregnant again 50-60 days after calving. Non-pregnant cows are known as open cows [129]. These open cows are 7.5 times more likely to be culled than pregnant ones [130] [131]. One of the most obvious signs that a herd is infected by *T. foetus* is an increase in the number of open cows and a reduced gestational period. The female bovine vaginal tract is home to a host of bacterial species [132] in comparison to the human reproductive tract which is 90% lactobacili, keeping the pH low (3-4.5) [133] [134]. One study [132] in cows found 792 different genera of bacteria between the uterus and the vagina and the abundance of species changed depending on the pregnancy status of the cows. The pH of the urogenital tract also changed, becoming lower when the cow was pregnant. In order to infect cows, *T. foetus* would have to be able to cope with these changes in environment. The bovine uterus also undergoes large changes after birth [132] [135] such as re-growing the uterine endometrium and shedding some of the uterine mucosa, which can put them at higher risk of disease caused by colonising bacteria. 80-100% of cattle were found to have contaminating bacteria in the lumen two weeks after giving birth, dropping to 40% after another week [135] due to clearance by the immune system. In order to infect the cow, *T. foetus* has to overcome the immune response and the vaginal mucus barrier [2].

1.4 Bovine Trichomoniasis

1.4.1 Clinical Signs of *T. foetus* Infections

The clinical signs of *T. foetus* infections differ depending on the animal host. In cats it causes diarrhoea, colitis, vomiting, weight loss and dehydration [16] [30]. The presence of the parasites also leads to histological features including a crypt epithelial cell hypertrophy and a loss of goblet cells [136].

In female cattle the clinical signs can be much more severe, causing vaginitis, infertility, endometriosis and in some cases pyometra (roughly in 2-8% of cases [137]) [138] [45] [139]. A *T. foetus* infection can also cause early embryonic death and abortion and the inflammation due to the *T. foetus* infection can cause placental oedema. Foetal pneumonia also occurs in roughly half of cases [140] and the abortion of the foetus can lead to longer breeding seasons. In bulls, *T. foetus* infections are largely asymptomatic, however, there can be small amounts of discharge present [139] [137].

In pigs *T. foetus* is primarily a commensal although it has been known in some cases to cause respiratory problems [141].

1.4.2 Prevalence and Economic Impact

Bovine trichomoniasis is a world-wide problem with cases of the disease being reported in the USA, Australia, Italy, Spain and Argentina amongst many others [142] [143]. The prevalence varies globally across a wide range of geographical locations. In Australia, for instance, bull infection rates in 1984 were 3.25%-34.6% [144] and infection rates in herds 11.5%-76.2%. This incidence led to a decrease in calf production of 3.6%-28.6%. In 1988 the estimated number of calves lost each year was 25,500 in the Victoria River district of Northern Australia [28], resulting in financial losses due to reduced milk and calf production estimated to be US $665 per infected cow. There is an estimated average financial reduction of 5% return per cow when the incidence of *T. foetus* infection reaches 20% among bulls in the herd. When the incidence of infection doubles then the financial reduction in turn increases to 35%. In the United States, infection has been detected in over 12 states [27], including Nevada where it is estimated that 5.8%-7.8% of bulls are infected; herd infections as high as 10.7% and it is thought that 40% of ranches have an infected bull [145]. In Floridian herds tested for *T. foetus*, the incidence was 6% overall (with a range of 0-27% within each herd). In Oklahoma, where calf production was found to be reduced by 18%, the projected financial loss due to *T. foetus* was US $7.3 million per year [138].

As bovine trichomoniasis is a global problem, the economic impacts are also very far reaching. It is estimated that the cost of this disease exceeds $650 million per annum in the US alone (Speer 1999) and so it is likely to reach into the billions worldwide, particularly as it often goes undiagnosed. There is a large reduction in the number of calves born and often the adults have to be culled if they test positive, both of these come at a large financial loss for the farmer. This is particularly evident for large scale producers, for example, on certain farms in Texas they could lose hundreds or thousands of infected cows. The income in Texas from cattle, stands at around $10 billion and is, therefore, the top agricultural commodity [146] [137].

There is also the cost associated with the feed and upkeep of infected animals that will not produce offspring, particularly for farms that only farm cows, e.g. for beef and milk, and have no other animals. This could lead to considerable financial losses. The potential loss of thousands of cattle per year could have a large impact on the economy, particularly in the beef and dairy industry. The amount of beef eaten in the world is estimated at 129 billion pounds (lb), [147] with countries

such as the United States and Brazil accounting for around 25 billion lbs and 16 billion lbs per year respectively. In the US the amount of beef consumed appears to be on the increase, potentially reaching over 28 billion lbs in the next few years. The lowest it has been over the past 20 years (in 2014) was still over 24 billion lbs so any significant loss of cattle would have a large impact on beef prices and would affect a large number of farmers. The amount of cows' milk produced globally is also very high: over 200 billion litres per annum and the amount of dairy products consumed in Europe is estimated at 46 billion metric tonnes [148]. In the US the production of milk is increasing from 160 billion lbs in 1999 to over 210 billion lbs in 2019. A decrease in dairy cow numbers, in a similar way to beef cows, could have a devastating impact on local economies, particularly in farming communities.

All figures gathered so far are from developed countries, such as the USA as the incidences rates of bovine trichomoniasis are not well reported in developing countries. It would seem likely that this disease is more wide spread in developed countries with large scale farming practises, in part due to the anthropogenic nature of the spread of this disease. Meaning that if there is one infected bull, the fact that this bull can be used as a stud for many females means that the disease will be spread more easily, even with the introduction of artificial insemination.

1.4.3 Diagnosis

The World Organisation for Animal Health has produced the OIE Terrestrial Manual which lists several ways to identify *T. foetus*. The samples from cattle are usually collected from the preputial cavity in bulls, the vagina for heifers or from aborted foetuses [16] [149] [150]. Samples from cats are faeces or rectal swabs and samples from pigs are nasal swabs. Preputial samples can include brushing, scraping or washing and all techniques are comparable with one another. Three consecutive weekly tests with a week break after sexual rest gives 95% or higher sensitivity [40]. The simplest and most widely used technique for identifying *T. foetus* is by microscopy the sensitivity of which is predicted to be 38-82% [16] . The parasites can be easily cultured in Diamond media and then examined using light microscopy [150] [16] . The culture is viable for up to 120 hours and at 37°C but it is easy to misdiagnose *T. foetus* if other trichomonads are present [151]. Dufernez (2007) [151] found 12 non-*T. foetus* isolates in bull preputial samples which could be mistaken for *T. foetus*. These included *Pentatrichomonas hominis*, *Pseudotrichomonas* species and *Tetratrichomonas* species.

There are also molecular methods for diagnosis. Polymerase Chain Reaction (PCR) can be performed, based on ribosomal RNA genes and can detect 1-10 *T.foetus* parasites per sample [38].

Real-time PCR is now commonly used [137]. In cattle, *T.foetus* often does not provoke a strong immune response. Therefore, there are relatively few serological tests for *T.foetus* antibodies [16]. There is an enzyme-linked immunosorbent assay (ELISA) based on a *T. foetus* antigen TF1.17 that detects vaginal IgA antibodies [152]. Rhyan (1995) [153] evaluated using immunohistochemistry using formalin fixed aborted foetal tissues. Monoclonal antibodies were used which detected *T.foetus* parasites in the sample whilst not labelling other trichomonad species. However, the sensitivity was very variable .

Diagnosis of infection by *T. foetus* may be compromised in several ways, including delays in shipping, contamination of samples and false positives. These factors can lead to unnecessary culling of healthy bulls with financial consequences for the farmer. The only accurate way to test is for infection is by constant sampling of the herd every two weeks, a procedure that is of relatively low cost of around $4 per animal [154], although this is not always practical, particularly in very large herds.

1.5 Pathogenicity

1.5.1 Mechanisms of Pathogenicity in *T. vaginalis*

There is more information on the mechanisms of pathogenicity in *T. vaginalis* as it has been more extensively studied. Some of these mechanisms of pathogenicity may represent reasonable hypotheses for *T. foetus*. Important factors for *T. vaginalis* pathogenesis include: adherence, hemolysis, phagocytosis, cell-detaching factors, the acquisition of host molecules and virus co-infection [155] [33].

Adhesion

Cysteine proteases were found to be necessary for adhesion to the host cell surface. Adhesion to the host cells causes degredation of the host-cell cytoskeleton [156]. During *T. vaginalis* infections lipophosphoglycan (LPG) along with fibronectin-binding proteins (FBP), GTP-binding proteins (GBP) and actinin, amongst others, are upregulated in the presence of the host and are thought to play key roles in adherence[157] [158]. Highly virulent phenotypes of *T. vaginalis* were found to bind more strongly to soy bean protein, than low virulence strains, providing further evidence

for their role in adherence and virulence. In 2008, it was found that *T. vaginalis* LPG binds to the human Galectin-1 and is inhibited by the presence of exogenously added *T. vaginalis* LPG [**?**], supporting a role for LPG in attachment. The exogeneous LPG also prevented binding of *T. vaginalis* to epithelial cells [159] [160]. This role for LPG could occur in other trichomonads and may be a mechanism to induce binding in different hosts.

Different *T. vaginalis* strains can have very different effects on vaginal cells [67] [161]. *T. vaginalis* can also bind to non-cellular structures, suggesting that the binding process is non-specific. There is a class of adhesion proteins (AP) that are though to be found on the parasite cell surface [162] [163]. They are involved in carbohydrate metabolism and many are found in the hydrogenosome. It is unclear whether these mediate specific or non-specific binding and what the true binding partners are. Membrane proteins, [20] serine proteases and cysteine proteases, are homologous some to aplicomplexan virulence and invasion proteins [164] and may have a role in host-cell degradation and lysis [20]. The BspA gene family encodes a very large family of leucine-rich-repeat containing proteins [67] that may mediate host-cell attachment in a similar way to some bacteria. these leucine-rich repeats are thought to act a recognition motifs for cellular binding. The BspA family have been found expressed on the bacterial cell and are involved in adhesion and invasion of host cells [165] and can trigger immune responses. The BspAs found in *T. vaginalis* contain a similar structure to those in *Bacteroides forsythus*, a bacterium which is known to be able to colonise oral cavities using BspA binding to fibronectin and fibrinogen [166]. Furthermore, bacteria lacking BspA genes were significantly less pathogenic than those with the genes [167].

Haemolysis and acquisition of host molecules

Beta haemolysin could be used to provide nutrients from the host cell to the parasites, for example fatty -acids, which *T. vaginalis* cannot produce itself [33]. In a study by de Carli (1996) [168] *T. foetus* was found to not possess haemolytic activity when exposed to a range of erythrocytes or several blood groups, however, *T. vaginalis* lysed cells from all of the human blood groups tested as well as the animal groups. The haemolytic activity was thought to be mediated by the cell attachment rather than by any secreted proteins or enzymes. When the parasites were pre-treated with Concanavalin A, the amount of haemolytic activity was reduced, further supporting the cell attachment hypothesis, potentially via sugars. Furthermore, in another study, when *T. vaginalis* cells were separated from blood cells by a permeable membrane no haemolysis occurred [169].
A glycoprotein was also found in *T. vaginalis* that mediates cell detachment from a monolayer

(found to be absent in *Pentatrichomonas hominis*) [155]. This detachment may allow the parasite to acquire nutrients or access further areas of the host since the monolayer has been breached. *T.vaginalis* also implements phagocytosis to acquire host cell nutrients [67]. This phagocytosis is also thought to potentially reduce the host immune response, leading to the asymptomatic nature of the disease [67].

Secreted proteins

Supernatant from media that has previously contained *T. vaginalis* cells has been found to have cytotoxic capabilities against host cells, suggesting a secreted virulence factor [33]. When the supernatant is added to a monolayer it can cause host-cell death [33]. *T. vaginalis* secretes exosomes that can bind and induce changes in the host cell [33], that may increase adherence, and proteases that degrade cytokines [67]. Cysteine proteases from *T. vaginalis* were often found in vaginal secretions taken from women infected with *T. vaginalis* and may have roles in cytotoxicity [67]. Certain cysteine proteases, such as Clan CA cysteine proteases, were also found to be produced in higher amounts in highly virulent strains [170]. When *T. vaginalis* strains showing low and high virulence [171] were compared there were several over expressed proteins in the high virulence strain, including cytoskeletal proteins and proteolytic enzymes. The low virulence strains showed no morphological changes when exposed to the vaginal epithelial cells (VECs), whereas the high virulence strains changed from trophozoites to a flattened amoeboid form [171] [172].

Virus co-infection

It has been thought that *T. vaginalis* can be infected with a double stranded RNA virus, known as *T. vaginalis* virus (TVV) and it may have an effect on *T. vaginalis* virulence, pathogenicity and drug resistance [173] [174] [175] [176]. It is thought that the prevalence of infected *T. vaginalis* with TVV could be as much as 50% [173] [174]. Interestingly, *Mycoplasma hominis* was also identified in over 80% of TVV positive isolates tested [174] and they may form an endosymbiotic relationship. Proteomes were analysed from both infected and non-infected *T. vaginalis* and 50 proteins were found to be altered between the two states [173]. In infected *T. vaginalis*, ribosomal proteins, isomerases, ABC transporters and adhesion proteins were upregulated, whereas heat shock proteins, malate dehydrogenase and fimbrin were downregulated [173]. The adhesion proteins, Ap33-1 and AP51-3 have been implicated in adhesion of *T. vaginalis* to host cells and their upregulation could lead to increased infectivity in TVV positive cells. It has also been seen that TVV can lead

to upregulation of cysteine proteinases [176], key enzymes involved in host cell degradation and monolayer destruction. Additionally, a surface immunogen, P270, was found to be upregulated in TVV positive isolates and P270 is thought to aid *T. vaginalis* in evading the host immune system. TVV positive isolates are also thought to have increased susceptibility to metronidazole, the usual trichomonad treatment, however, the mechanism by which this susceptibility increases is unknown [175] [176].

1.5.2 Mechanisms of Pathogenicity in *T. foetus*

During infection *T. foetus* creates close cell contact and can be cytotoxic towards the host cells in some cases. [38].
In vitro, *T. foetus* can attach to bovine vaginal cells in addition to MDCKs and CHO cells [38] and the parasite has been shown to have a high cytotoxicity to Hela cells. It is thought to attach to host cells via 'filopodia-like protrusions' [38] and this adhesion could be a key stage in pathogenicity. The parasite attaches to the mucosa, such as the gastrointestinal tract or urogenital tract, however, there is evidence that it can also penetrate the mucosa of foetuses [153] and invade neighbouring tissues.

Adhesion of cells was found to be enhanced by laminin and inhibited by anti-laminin suggesting that laminin could be important for cell adhesion. Additionally, *T. foetus* and *T. suis* were found to bind well to porcine mucosa, mediated by lectin-like binding. The presence of sialic acid on the host cells improved binding [38].

It has also been thought that *T. foetus* could cause damage to the female urogenital tract by affecting the usual vaginal flora, such as lactobacillus, thereby, changing the pH and other conditions [38]. When bovine sperm was incubated alongside *T. foetus* cells, it was found that the parasite damaged the sperm and caused agglutination [177], whilst also decreasing sperm motility. The overall viability of the sperm did not seem to be affected [177].

It has been found that *T. foetus* can also secrete many different proteases, especially cysteine proteases, into the media. These proteases can cleave host-cell proteins such as fibronectin and fibrinogen [38] [177] which could damage the host mucosa. Sialadase was also found to be secreted and it has been postulated that the cleaved host sialic acid could be a nutrient source for *T. foetus* [38]. A similar sialadase has also been found in the related trichomonad *T. mobilensis*.

1.5.3 Pathology

Only 100-200 motile cells are required for infection and cows can usually eliminate the parasite after 2 months [178].The bull often shows signs of inflammation for up to 2 weeks before becoming an asymptomatic carrier. In cattle, it is thought that the first step in colonisation is the attachment to the mucosal euterovaginal epithelium; likewise in felines, there is attachment to the intestinal epithelial monolayers. During infection, *T. foetus* is only found on mucosal surfaces and the main site of infection in heifers is the cervix. However, the effector mechanisms during infection have not, as yet, been identified. In males, *T. foetus* does not localise in the epithelium, rather in the actual secretions of the epithelial lining [26]. The progress of infection is rapid, with all regions of the reproductive system being found colonised within 2 weeks [28]. Infection by *T. foetus* causes inflammation in uterine tubes and endometrium and *T. foetus* can, in some cases, cause permanent sterility [38]. It takes between 2-6 months for a cow uterus to fully recover from an abortion [28]and the cows can become reinfected at the next breeding session. The parasite cells can adhere directly with the vaginal mucosal epithelium [179] before occupying the uterus. The *T. foetus* cells are then thought to change morphology from trophozoite to pseudocyst form.

1.6 Immunology

1.6.1 Immunity in Cattle

In cattle there is little evidence for acquired immunity. In males, although some parasite specific immunoglobulins have been identified by some studies in small amounts [5] [180]. In females, *T. foetus* specific IgA and IgG responses are recorded [181] [5] in vaginal secretions and urine. The IgA mediates parasite immobilisation and agglutination, whilst IgG is responsible for complement recruitment. When bovine immune sera were examined, there was complement-mediated killing of *T. foetus* [38]. There are different responses dependent on whether an infection is natural, or experimental and the parasites have been administered to the cattle.

Immune responses have been describes in cattle following experimental infections [182]. The antibody titre in cattle infected with the Belfast strain of *T. foetus* [5] increased from 32 before infection to 512 in heifers and 128 in bulls four months post infection. The heifers showed antibodies in vaginal mucus and became parasite-free after four months. The serum antibody response was predominantly IgG2, with minor contributions from IgA and IgG1; antibodies in vaginal mucus were IgG1. No antibodies were seen in the preputial secretions of the bulls. *T. foetus* does not

penetrate the preputial epithelium and so its antigen levels may not be high enough to provoke an immune response.

In another experiment, Corbeil *et al.* [181] examined the immune response in the female bovine reproductive tract and compared protective efficacy in animals with genital responses of predominantly IgG or IgA with a purified surface antigen (TF1.17) of *T. foetus*. Ten cattle were used as unvaccinated controls. The other cattle were immunised twice with the TF1.17 antigen and then either subcutaneous (Group S) or intravaginally (Group S/V) killed *T. foetus* was administered. Three weeks after the last immunisation, animals from all groups were challenged with *T. foetus* parasites. Samples of vaginal mucus, serum and urine were collected. Very low IgA responses were seen in vaginal secretions for all groups, however, group S/V had the highest uterine responses. Both treated groups, S and S/V had members that shows no evidence of *T. foetus* infection, 35% and 20% respectively. Both groups also cleared the remaining infection faster than the controls and there was found to be no statistical difference between clearance rate between the two treated groups. Serum IgG responses were high in both treated groups, though group S remained higher throughout the experiment and vaginal IgA was higher in the S/V group.

1.6.2 Potential Mechanisms of Immune Evasion

Tritrichomonas foetus has several ways in which to evade the bovine immune system. It may colonise the vaginal tract where a relatively low concentration of complement is found, thus providing protection from the cow's immune system [38]. The cysteine proteases that it secretes can, *in vivo*, cleave IgG1 and IgG2, the main antibodies mediating killing of the parasite. IgA does not seem to be lysed to the same degree as IgG; however IgA is involved in preventing attachment rather than mediating killing [38]. There is evidence that the parasite can internalise the antibodies, preventing them from binding effector mediators such as complement or granulocytes [38]. The idea of antigenic variation has also been suggested as in one study there was a variation in expression of TF190 [38] the surface LPG undergoes epitope variation, however, the major portion is conserved [177] [183]. Furthermore, if *T. foetus* sheds its antigens, such as TF1.17, from the surface, antibodies could bind to these away from the main cell, potentially reducing cell killing [177].

1.7 Prevention and Treatment

As *T.foetus* infections can cause considerable issues to farmers and vets, both financial and in terms of animal welfare, preventative measures and treatments have been produced.

1.7.1 Prevention-Biosecurity measures

The most successful method for controlling *T. foetus* infections is considered to be biosecurity practices and routine testing of animals [139] [5].

Early breeding of heifers

Making heifers become open earlier in the breeding season so they can then be mated with an infected bull. They would then become temporarily naturally immune to *T. foetus* and could be allowed to breed later in the season and could carry the calf to term. This comes with costs, principally, the amount of cow feed to allow the heifers to enter the breeding season early.

Using virgin bulls

Only breeding using virgin bulls that have never been exposed to *T. foetus*. The use of younger bulls also decreases the chances of infection compared to older ones (over the age of four) as they are naturally more resistant to infection [40]

Culling infected animals

Culling has a large financial impact to the farmer or owner. Culling infected animals will prevent further spread of the disease. Open heifers and those that have been deemed to be pregnant but have then not carried to term are likely to be a reservoir of infection and could promote an infection at the next breeding season. there is also the danger of herds becoming infected again from neighbouring farms.

Artificial insemination

Using only known uninfected bulls is a way to prevent a whole herd being infected by one infected male. The rise of artificial insemination (AI) could be seen as a key factor in reducing infection

rates in developed countries. Most cases of *T. foetus* infection come from farms still using natural service rather than AI [5]. The proportions of dairy farms, in North America, using natural service in 2005 was 40% [5]. However, there have been some cases where AI has not been successful in preventing the spread of *T.foetus*. Cases of poor practise have led to increased transmission, for example, failing to change gloves when manually checking cervico-vaginal mucus from cows or failing to screen semen for *T. foetus* contamination [5].

Replacement animals

Any new bulls or heifers brought into the herd should be tested for infection, males brought in should be virgin bulls or under the age of 4. Any cows that do not become pregnant or carry to term should be removed from the herd. Dairy cows also have year-round breeding seasons, unlike those in beef herds. This allows the herd as a whole to promote the spread of *T. foetus* all year round, rather than at specific times and can make it more difficult to assess if the pregnancy rate of the herd is decreasing due to infections.

Monitoring of infection

Should be done prophylactically within herds. Preputial scraping from bulls over the age of 3 should be taken each year before breeding. Samples from aborted foetuses and pyometra should be taken and examined for *T. foetus* parasites. If there is a suspicion that heifers in a herd are infected, cervical mucus samples should be taken, cultured and examined for the parasites.

1.7.2 Treatment

As mentioned previously, imidiazole-ring compounds are no longer used as a *T. foetus* treatment and resistance to them is already being seen in parasite populations [184]. Therefore, good biosecurity practises are currently being seen as a more effective way to control disease spread. Historically, trypaflavine ointment was applied to the penis of the bull [139], however, this required several treatments, was expensive and difficult to apply and is very rarely used now. Dimetridazole given to bulls orally or intravenously can be used, however, it has been known to cause them to collapse. Bulls also have to be treated with penicillin too as some bacterial infections can prevent dimetridazole from working [139]. These treatments do not prevent reinfection. For infected animals it is usually those that are highly genetically valuable that are treated and saved whilst the others are culled.

1.7.3 Current *T. foetus* Vaccines and Vaccine Attempts

Currently there is no approved chemotherapy for *T.foetus* in the US as the previous treatments (imidazole ring compounds) for infection have been removed from the market [1]. This is due to the numerous side effects particularly, those affecting pregnant animals. The only current vaccine available is Trichguard, a killed vaccine, but it neither cures infected cows nor prevents re-infection. It does however, reduce the shedding of the parasite by heifers, decrease rates of spontaneous abortion and as a consequence lead to an increase in calving rates from 60-70% to 80-85% [185]. Trichguard does not appear to reduce bull associated infections nor is it licensed for use in bulls [23][186].

Fuchs *et al.* [187] isolated the B1 strain of *T. foetus* and inactivated the parasites using formaldehyde. They then inoculated 74 heifers using a variety of adjuvants. No clinical signs of infection were seen. IgG increased after the first immunisation in all samples, but IgA remained low until the challenge with live *T. foetus* parasites [187]. All animals showed colonisation of the urogential tract and two of the twelve unvaccinated animals were still infected after 105 days. No animals became infected in the vaccine groups, apart from one animal when the vaccine was applied with an oil-adjuvant. The average clearance time was 55 days for the immunised animals and 69 days for the non-immunised animals [187].

TrichGuard is a commercially available killed whole cell vaccine [150]. Edmonson (2017) [150] tested the efficacy of Trichguard in an experiment using 40 heifers and another 20 were used as controls, receiving two sham vaccinations of a saline solution. 60 days after vaccination the heifers were inoculated with an isolate of *T.foetus*. Vaccination with Trichguard increased the pregnancy rate from 70 to 95%; however, this was one study on a small number of cows and it has not been shown how well the vaccine works long term, i.e. over a number of years [150]. The average clearance time for control animals was 79.15 days and for vaccinated animals was 63.9 days, a non-significant reduction. The vaccine appeared to improve conception rates but this is contradicted by other studies [188] [189]. If a vaccine does not confer long lasting immunity and must be reapplied every breeding season then it is not practical for large scale farms or cattle drives where there could be hundreds or thousands of cows to vaccinate every few months.

As there is no vaccine that can prevent re-infection the current modes of control are mainly directed towards prevention rather than cure. Hudson (1993) [1], compared *T. foetus* vaccines based on whole cells with those composed of cell membrane fragments. Animals vaccinated with whole cells had higher clearance rates than cows which were vaccinated with cell membrane fragments.

Furthermore, cows that were vaccinated twice had higher *T. foetus* clearance rates than those that were vaccinated only once. All cases showed an improved response over those in the control group that had not been vaccinated.

However, cattle have been found to have a limited protection against *T.foetus* in the form of infection induced acquired immunity. Adhesion of the parasite to the host cells appears to play a central role in the cytopathic mechanism of host cell damage [190]. Shaia *et al.* [191] identified an adhesion molecule Tf190 on the surface of *T.foetus*. They also studied the humeral responses of cattle to immunisation with Tf190 and detected a strong systemic response. They found that immunisation with Tf190 decreased the infection in immunised animals compared with those not immunised. Furthermore, a parasite-specific antibody was detected in the cervical mucus of the immunised animals when *T.foetus* was inoculated into the vagina. Chromatography showed that different epitopes were expressed in different strains of *T.foetus* but the Tf190 adhesin complex as a whole is widespread throughout strains. These differences in expression appear to be based on the geographic isolates of *T.foetus*.

When bulls were vaccinated using killed whole cells, a curative effect failed to be demonstrated [192]. There was a reduction in duration of infection in vaccinated heifers when challenged by bulls infected with same *T. foetus* isolate however, no significant practical advantage was seen. One out of twelve heifers ended up becoming pregnant in the vaccinated group compared to two out of twelve in the control group.

In another study by Rhyan (1999) [180]; 24 bulls that had history of infection were sampled- *T. foetus* was identified in 15 of the bulls using monoclonal antibodies 34.7C4.4 and TF1.15 against *T. foetus* surface antigens. Preputal secretions from the bulls were tested and were found to contain specific isotypes and each isotype response was much greater in infected bulls than controls. Within the samples, IgG1 responses were the highest, IgM and IgA were equal and IgG2 responses were low. The TF1.15 antigen appears to be widely conserved between *T. foetus* strains and when anitbodies were used there appeared to be strong labelling of the myelin shealth by the TF1.15 antigen [180]. The antigens with serological specificity seemed to be on the outer surface of the parasite and possess polysaccharide and protein components [144]. TF1.15 was shown to be highly glycosylated with a glycolipid moiety [180].

Clark (1984) [144] conversely, found that injections of killed *T. foetus* cells could prevent and eliminate genital infection in bulls. Their results suggested that bulls can be immunised, 3 of 4 of bulls vaccinated remained uninfected. This has to be taken with caution as the sample size of 4 bulls

suggests that at least 25% of bulls would still be infected if this was implemented as a vaccine. It is likely that the non-infected bulls' natural immunity played a part. Larger trials using the same methods would need to be used to give a true outcome. Additionally, genome-wide association studies (GWAS) could be used to identify whether certain bulls have natural immunity to T. foetus and whether there are any genetic markers to identify this (Ku 2010, Hart 2015). It may also be possible to identify gene markers of the most susceptible animals as, as previously stated, certain breeds of cattle have higher rates of infection than others [26].

Baltzell *et al.* [193] also looked at numerous studies relating to the use of a whole cell killed vaccine. They looked at several parameters, including: the incidence and duration of infection in both bulls and heifers; the percentage of pregnancies and abortion risk and the overall efficacy of the vaccine to clear the infection. The Grading of Recommendations, Assessment, Development and Evaluations (GRADE) quality of evidence in most of the studies was deemed low, in some cases due to publication bias. There were also issues with small sample sizes and non-randomisation of the sample animals. The conclusion that had to be drawn from this meta-analysis was that there was little or no evidence that the killed-whole cell *Tritrichomonas foetus* reduced infections or risk to open heifers or bulls.

1.8 Vaccinology

1.8.1 Classical Vaccinology

Classical vaccinology involves the administration of live attenuated or killed cells, cell fractions, native or recombinant antigens to ellicit an immune response and aquired immunity. For example, with influenza virus [194] [195], where the vaccines are made from the hemagglutinin (HA) and neuraminidase (NA). The methods of classical vaccinology have often proved effective, whole cell vaccines such as the inactivated polio vaccine and live attenuated vaccines like the tuberculosis BCG vaccine all elicit strong immunity. However, in cases of cell fraction vaccines, finding the fragments required to promote an immune response can be a slow process as each potential antigen must be looked at in turn. In addition, to locate genes and determine the effects they have on the organism using a forward genetics approach mutant phenotypes are typically made, and the effect it has on the organism is tested. Since the advent of genome sequencing it has been possible to examine the antigenic repertoire as a whole. As all antigens expressed by a pathogen are encoded

in its genome, it follows that it is possible to use the genome sequence to screen possible antigens for desirable properties in concert with, or in place of, forward genetic approaches [196].

1.8.2 Reverse Vaccinology

Reverse vaccinology involves designing a vaccine from the genome sequence information rather than having to grow the organisms and use fractions to promote or measure a response [197] and came into fruition by the advent of whole genome sequencing [198]. The evolution of this field has been greatly helped by improvements in computing software and bioinformatic analyses. In a reverse vaccinology approach, potential antigens are first found *in silico*, i.e. using computational approaches to identify structural properties in predicted antigens and are then verified using experimental methods. In many cases this is by screening the putative antigens by way of an assay or protein array to determine whether they promote an immune response and also by using animal models.

Reverse vaccinology does not just use whole genome sequencing alone but in combination with transcriptomic and proteomic approaches. Transcriptomic analysis using RNA-Seq can be used to assist in the identification of proteins from open reading frames (ORFs) and the characterisation of the expression of potential antigens and vaccine candidates. This can be particularly useful in finding start and stop sites and splice sites (in the case of eukaryotes). Proteomics can be used to validate the results from the *in silico* predictions. In the case of *Salmonella typimurium*, mass spectrometry- based proteomics was used as an accompaniment to the *in silico* predictions, finding several novel genes and confirming 40% of the ORFS already found [199].

1.8.3 Reverse Vaccinology Examples and Successes

The first vaccine created by a reverse vaccinology approach was that of the MenB vaccine [200] [201]. This was designed to protect against sepsis and meningitis caused by serogroup B strains of *Neisseria meningitidis* (MenB). The whole genome was analysed and proteins that localised to the cell surface were identified. Over 2000 predicted genes were identified, from which 350 recombinant antigens were created, 28 of which were selected for testing [201]. These were tested for immunogenicity against a strain collection of *Neisseria* species so that a 'universal' vaccine could be created, covering as many strains as possible. The final vaccine contained four antigens, including a heparin binding factor and an outer wall vesicle protein [201]. Using multiple antigens within a vaccine has been found to be more effective and to confer better protection than just one antigen [202] [203]. The effectiveness of this vaccine has been estimated at 83% [201].

Starting with the whole genome of *Pseudomonas aeruginosa* Bianconi *et al.* [202] identified 52 potential antigens using bioinformatics approaches. The candidates found were narrowed down from over 5500 predicted protein sequences and were conserved across strains. A number of outer membrane proteins and virulence factors were among these. A shortlist of 10 potential antigens was tested in mice and were found to elicit an immune response, however, the true biological function of these antigens is unknown.

It is also possible to base antigen design on potential structures found by computational methods and this was the method adopted when designing an HIV vaccine. A vaccine for the respiratory syncytial virus (RSV) was based on an epitope-focused design. However, with this method of vaccine design there can be the problem of naturally occurring epitope variants [200]. Reverse vaccinology also allows for studies into antigen function and several discoveries were made whilst using this approach, such as the discovery of the H binding factor, which led to the idea of species specificity in meningococcus; and the discovery of pili in Gram positive organisms [200]. For the *Helicobacter pylori* vaccine, the proteome and also comparative genomics were used [204]. The 3D structure and protein-protein interactions were found and Naz *et al.* (2015) then screened the exoproteome and secretome of samples for non-human homologous proteins and epitopes that could bind B cells and T cells.

To design a *Streptococcus agalactiae* vaccine [205] eight genomes were used as there are represented differences between strains. This allowed a much higher number of genes to be screened. A 'core' genome was identified that cotained genes found in all strains as well as a 'dispensable' genome containing genes that are only present in one or more of the strains. When these were screened *in silico* a large number of cell surface expressed core genes were found. Some of these were then expressed in *E. coli* and found to confer some immunity in mice. Several potential immunogenic genes were also found in the dispensable genome set. There were *E.coli* [205] antigens found in a pathogenic strain that did not feature in a non-pathogenic strain. Proteins were also found that provided protection in mice in recombinant form.

Applications to Parasites

There have been many attempts at producing malaria vaccines, based on various approaches such as sporozoite-specific antigen subunits, killed whole cells, as well as peptide and toxin based vaccines [206]. Whole organism sporozoite vaccines have been used to confer immunity and have been used to guide the vaccine production process for *Plasmodium*. Most malaria vaccines only target

one life stage of the parasite, usually the pre-erythrocyte stage [206]. Duffy (2012) [207] identified a pre-erythrocytic recombinant vaccine for malaria that does not cause infection but still confers protection against *Plasmodium*. Many attempts to ecauate cell-surface proteins as vaccine candidates found they were poorly immunogenic. Some blood stage vaccines target parasite antigens that are expressed on the surface of the host red blood cells. However, they are often highly polymorphic and not conserved across strains [208].

Due to difficulties with classical vaccinology approaches, reverse vaccinology has now also been applied to malaria vaccine design. This involved looking at the transcriptomes and proteomes of several different *Plasmodium* lifestages and identified several potential vaccine targets, including cell surface proteins [208]. This also had the advantage of identifying which transcripts were stage specific. The most effective vaccine would be one that targets multiple lifestages and could be deployed widely. Furthermore, conserved genes and homologs between species, such as reticulocyte homolog (Rh) proteins, have been found which could be used one day [208].

In the case of human *Leishmania*, there is still currently no vaccine, one reason being the challenge of vaccinating against multiple life stages [209]. Initially live attenuated vaccines were used. Second generation *Leishmania* vaccines include recombinant vaccines and have been licensed in some countries [209]. Potential vaccines and candidates could be based on secreted proteins or could include intracellular proteins like heat-shock proteins, histones, ribosomal proteins or cysteine proteases. *Leishmania* cysteine proteases are known to produce a higher IgG response than others intracellular proteins. *T. foetus* produces cysteine proteases and so if these are successful candidates in leishmania they also may have potential as *T. foetus* vaccine candidates.

Leishmania, like *Plasmodium*, has multiple life stages and the cell surface has been a target for possible vaccines. Promastigotes have a cell surface covered with glycosylated proteins which is potentially similar to *T. foetus* cell surface and many GPI anchors. LPG is found on the *Leishmania* cell surface and is used to protect the parasite against harsh environmental conditions. LPG is also found on the *T. vaginalis* [20] cell surface and is likely also found on *T. foetus* and could be another potential target. However, as LPG is formed using phosphodiester bonds and these are an example of a post translational modification (PTM), they will not been seen in the genome sequence itself. In order for it to be classified, a likely gene would need to be expressed in a eukaryote, such as yeast, before being tested for immunogenicity.

1.8.4 Comparing Classical Vaccinology and Reverse Vaccinology

Both classical vaccinology and reverse vaccinology have been used to produce successful working vaccines. The main difference between the two is that reverse vaccinology is not hypothesis driven and is a 'top down' approach [198] where as classical vaccinology proceeds from initial hypotheses. Reverse vaccinology also relies on initial computational methods which predict likely antigens [198].

With reverse vaccinology all antigens encoded by genome even non-immunogenic and from organisms that are cannot be cultured and the most highly conserved antigens can be found, rather than just 10-20 [210]. Non-structural proteins and non-immunogenic antigens can be found in addition to those that are not expressed *in vitro*. It is possible to compare the sequences of all the putative antigens before moving *in vitro*. This allows for the identification of homologous antigens in other species, for example, humans or cattle. The vaccine candidates chosen would need no homology to the host so that the vaccine would not damage to host itself. Comparing the sequences also allows for comparisons between the antigens themselves to check for antigenic variability between strains. However, with reverse vaccinology only protein antigens can be screened and found rather than glycolipids and polysaccharides which can be identified and used with the classical vaccinology approach.

1.9 Key *T. foetus* Vaccine Candidate Criteria

In order to identify vaccine candidates to later test, we used several criteria that our candidates were to adhere to. These were:

1) Cell surface expressed [211] [212] [213] [214]
2) Antigenically invariant across the population [211] [215]
3) *T. foetus* specific [216] [217]
4) Immunogenic [218].

As the cell surface is the part of the parasite that comes into contact with the host, many of these proteins are likely to confer an immune response [213] and, ideally, are required for parasite survival [217] [219]. Therefore, these are may be good vaccine candidate that will help promote an immunogenic response. Limiting the vaccine candidates to cell-surface proteins also reduces the

time and cost associated with finding candidates as they can be predicted computationally [220]. In order to find cell surface expressed proteins certain motifs were looked for: namely signal peptides, trans-membrane helices and GPI-anchors, all which suggest cell surface localisation (Figure 1.6).

To promote a strong response to the *T. foetus* parasites, proteins were examined for trichomonad specificity. This is important as the vaccine must confer a response to the pathogen and not the host. We also did not include any putative bacterial proteins or LGTs to further reduce our vaccine candidate number. When BLAST searches were performed on the *T. foetus* genes, any genes that either have no BLAST hit or have a hit to a sequenced trichomonad genome were then examined in more detail to ascertain whether they fulfill the other criteria. Orthofinder [221] [222] will also be used to identify similarities between *T. foetus* and *T. vaginalis* genes, producing probable gene families. This allows likely gene function to be identified in cases where there were no BLAST hits.

Samples of *T. foetus* from different sources (different strains) will be used to make sure the proteins of interest are antigenically invariant across populations, therefore, a vaccine against one strain will work against all others rather than different vaccines being needed for different regions.
For the antigens to be successful vaccine candidates they must confer an immunogenic response, thereby allowing the host to acquire long-lasting immunity.

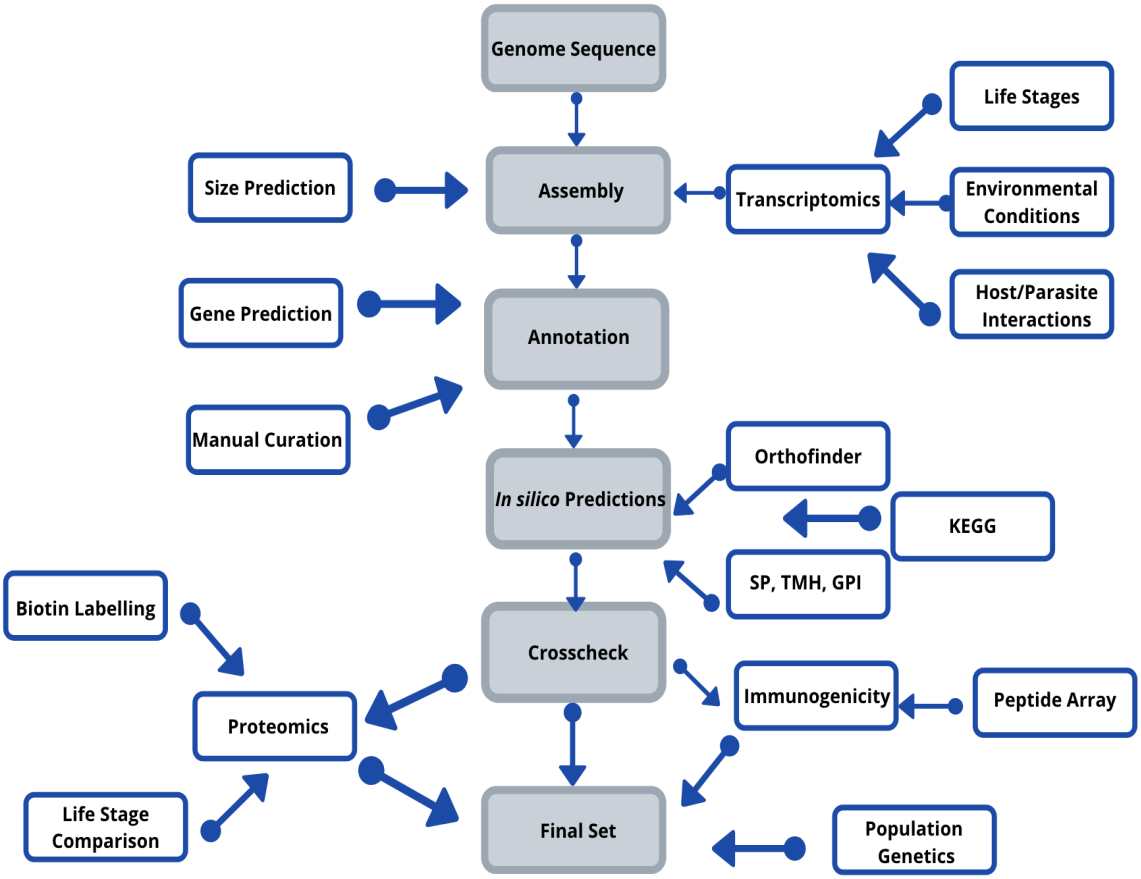

Figure 1.6: The stages leading to a vaccine candidate set. Showing all stages from obtaining the genome sequence to the final candidate set with the different components leading to the completion of each stage. Where SP refers to signal peptide, TMH refers to transmembrane helix, GPI refers to Glycosylphosphatidylinositol anchor and KEGG refers to Kyoto Encyclopaedia of Genes and Genomes.

1.10 Aims of the Thesis

The overall aim of this project was to identify candidates for a bovine trichomoniasis vaccine using a reverse vaccinology approach, by producing an annotated *T. foetus* genome sequence and evaluating all possible antigen genes found within it for their immunogenic potential.

A *T. foetus* genome has been sequenced on the PacBio platform and then assembled by SMRT portal [223] [224]. The newly assembled genome was then annotated using gene predictive software in conjunction with transcriptomic and proteomic data. The resulting genome was 147Mb in size, with 2,776 contigs and an N50 of 82,179. This appears to be comparable with the *T. vaginalis* genome of 160Mb. Additionally, the genome produced has a similar BUSCO [225] [226] score to that of the *T. vaginalis* genome, showing completeness.

As stated previously, potential vaccine candidates were identified from antigen proteins that fulfilled several criteria, they were:

1) Parasite-specific [216] [217]

2) Cell-surface expressed [211] [212] [213] [214]

3) Naturally immunogenic [218]

4) Consistently expressed

5) Upregulated during host interactions

6) Structurally invariant across the parasite population [211] [215].

To evaluate these criteria we:

1) Identified *T. foetus* genes containing structural motifs that predict surface expression

2) Validated putative cell-surface expression using cell-surface proteomics and biotin labelling

3) Screened parasite peptide microarrays with infected bovine serum

4) Quantified antigen gene expression using transcriptomic assays *in vivo*

5) Quantified structural variation in antigen genes sequencing DNA isolated from several different *T. foetus* strains and identified SNPs

This project produced validated vaccine candidates as a first step in producing a viable vaccine against bovine trichomoniasis, facilitating clinical trials into the efficacy of vaccination as an effective strategy to eradicate the disease.

Chapter 2

The Genome and Transcriptome of *T. foetus*

This chapter describes the assembly and annotation of creation of a *Tritrichomonas foetus* genome from PacBio DNA reads. It follows the steps from assembling the reads to the identification of the open reading frames (ORFs), subsequent annotation of those ORFs and mapping of putative gene functions. It then compares the genome annotation to those already produced for *T. foetus* by Benchimol [227] and Carlton [228] and examines at the merits of each including how the number of putative genes differs between them. This chapter will also look at which annotated proteins are predicted by bioinformatic methods to be on the *Tritrichomonas foetus* cell surface, a key feature we are looking for in our vaccine candidates. Finally, to improve and enhance the genome annotation, gene expression is examined in this chapter under various developmental and environmental conditions.

2.1 Introduction

2.1.1 Genome Sequencing

More genomes are being sequenced now than ever before. This is mainly due to falling costs in sequencing technology, for example $50 for a bacterial genome, and also the availability of in-house sequencing platforms. According to Moore's law [229] the cost of genome sequencing will get lower as time goes on. Many genome technologies have been created in recent years, such as PacBio sequencing and Nanopore MinION sequencing, each having their own benefits depending on the composition of the genome itself and conditions in which the DNA has been produced.

Illumina has long been used as a sequencing platform, producing short reads (up to 250bp) and the machines are constantly being updated and improved, allowing millions of reads to be produced quickly and relatively cheaply. In contrast, PacBio sequencing, also known as single molecule real time sequencing (SMRT), produces long reads (up to 20kb) [224]. These reads are better for spanning repetitive regions, for example, the large repeats in *T. vaginalis* genome. This is beneficial as key genes could be found in these regions and makes it more suitable for the sequencing of this type of genome compared to Illumina. As the *T. foetus* genome is likely to have a similar makeup to *T. vaginalis* it would be reasonable to assume that it too would have a large number of repetitive regions.

2.1.2 Gene Prediction Software

Predicting genes in *de novo* genomes, especially in eukaryotes if they are not closely related to a model organism, can pose many problems. There may be non-standard codon usage, splice sites and regulatory signals which would be difficult to identify. Therefore, specialised software would need to be used and 'trained' so that ORFs within a particular genome can be predicted. In order to get the most accurate results, several programs can be used to obtain a consensus. There are many different software that employ *ab initio* techniques, meaning they predict the gene models from the genome sequences alone [230]. In the absence of any experimental data alongside the genome sequence, these have been found, to be inaccurate[230]. *De novo* gene finders include SNAP and AUGUSTUS [231]. Due to the inaccuracies, that appear to be common to *de novo* gene finders, using other software that requires experimental data, for example BRAKER which employs transcriptomic data, and manually curating the resulting ORFs can produce a more accurate consensus.

AUGUSTUS

AUGUSTUS is based on a general Hidden Markov Model (HMM) and submodels and predicts the most probable distributions for introns and exons [231] in addition to predicting translation starts and ends. The performance of AUGUSTUS can be improved by integrating BLAST or sequence alignments. AUGUSTUS also has the advantage of being able to identify alternative splicing, which is key when annotating an eukaryotic genome. When three *de novo* gene finders (AUGUSTUS, SNAP and GLIMMERHMM) were compared using the *Toxoplasma gondii* genome, AUGUSTUS was found to have the highest specificity and sensitivity [230] and also appeared to have the highest accuracy. Additionally, as these gene finders were tested on a eukaryotic protist

parasite genome, it is reasonable to assume that AUGUSTUS would also have high sensitivity and accuracy when used to identify genes in the *T. foetus* genome.

SNAP

Semi-HMM-based Nucleic Acid Parser (SNAP) uses HMMs in a similar way to AUGUSTUS [232]. It is known to be accurate in eukaryotic genomes, such as *Drosophilia* and *C. elegans* when trained and has been found to give fast and accurate gene models even when the training set contains minimal promoter UTRs [232]. It also models each strand separately and so allows introns on different strands to overlap each other.

BRAKER

BRAKER [233] uses RNA-seq data to annotate genomes and combines aspects of AUGUSTUS [231] and Genemark-ET [234]. Both programs are trained by BRAKER before being applied to the genome [235]. Genemark-ET creates the initial *ab initio* gene structures and AUGUSTUS uses those structures to train and then integrates the RNA-Seq data. BRAKER is also interesting because it integrates RNA-Seq data in the gene models to provide further evidence for likely splice sites and stop and start codons.

2.1.3 *T. foetus* Genomes

In 2017, two draft *T. foetus* sequences were produced [227] [228] and then in 2020, [120] samples were taken in order to gain an understanding of the prevalence of *T. foetus* across Europe in cows, pigs and cats, producing three more genomes from bovine, feline, porcine strains of *T. foetus* respectively [120]. These genome sequences were all produced using Illumina sequencing and short DNA reads.

The Benchimol *et al.* genome [227] was produced using the feline strain 'K' as this strain has been used in several other *T. foetus* studies [236]. The study was undertaken to develop *T. foetus* as a model organism for other trichomonads and parabasalids. Additionally, it was sequenced in order to allow a comparison between *T. foetus* and *T. vaginalis* as they appear to be similar in size and have similarly large amounts of repetitive regions [227]. It was assembled using AllPaths-LG [237] to produce 3,730 contigs in 1,573 scaffolds to create an assembly of 68Mb. Allpaths-LG estimated the total genome size to be 161Mb, which is over twice as large as the actual genome assembly and more in keeping with Zubakova *et al.* (2008) [114] predictions and the size of the *T. vaginalis*

genome. This suggests that the Benchimol [227] genome assembly is only half the correct size, possibly due to collapsing large repetitive regions. The percentage of repetitive sequences was estimated to be 62%, which is comparable to the *T. vaginalis* genome. The System for Automated Bacterial Integrated Annotation (SABIA) [238] was used for gene finding and this identified 7,856 proteins and 17,497 hypotheticals in the Benchimol *T. foetus* genome assembly [227]. However, SABIA is designed for bacterial genomes and therefore it is possible that genes or splice sites in this eukaryotic genome may have been misplaced due to differences in gene structure.

Carlton *et al.* [239], in a similar way to Benchimol [227], sequenced *T. foetus* in order to compare with *T. vaginalis*. In particular, they wanted to compare conservation of biochemical and metabolic pathways between the two species. Carlton *et al.* assembled a *T. foetus* genome sequence using Velvet [240] [241], producing 194,695 contigs with an assembly size of 67Mb and N50 of 2,054. This sequence contained 28,362 ORFs with orthologs in the *T. vaginalis* genome [228]. This comparison was particularly important as *T. vaginalis* is becoming resistant to one of the FDA approved treatment drugs, metronidazole (Mz) and there are now cases of resistance in *T. foetus* . The KV1 strain of *T. foetus*, along with two Mz resistant strains (aerobic resistance only and both aerobic and anaerobic resistance) were sequenced. BLASTX of *T. foetus* ORFs against *T. vaginalis* showed 28,600 orthologues with 183 SNPs in total in these orthologues. Thirteen of these SNPs were present in both of the *T. foetus* resistant lines, suggesting that they may be associated with a gene resistance phenotype. The results showed that there may be a shared pathway of Mz resistance in both species.

Westrop *et al.* [242] also compared metabolic profiles of *T. foetus* and *T. vaginalis*, particularly with respect to Mz resistance. Several differences in the metabolic pathways were seen between the two species and also within the same species depending on whether the strain was Mz resistant. For example, the *T. vaginalis* products for glucose metabolism include; glycerol, ethanol and lactate. In *T. foetus* they are largely the same, however, *T. foetus* produces succinate rather than lactate [242]. In Mz resistant strains of both parasites the metabolism is shifted. In the case of *T. vaginalis*, it is shifted to a much higher output of lactate in but in *T. foetus* it is shifted to a higher output of ethanol. It was also noted that due to its repetitive regions and predicted large genome size, the *T. foetus* genome could likely have undergone massive gene expansion in a similar way to *T. vaginalis*. In the case of Dabrowska [16], the genomes produced were mapped to the Benchimol strain K sequence using Bowtie2 [243] and SNPs were identified using the GATK pipeline [244]. No estimation of size was stated, nor how many genes found. However, the study by Dabrowska *et al.*'s data does provide evidence that the porcine and bovine strains of *T. foetus* are genetically much less distinct from one another than they are from the feline strain. Other studies, for example,

Slapeta *et al.* (2012) identified differences between cysteine proteases in feline and bovine strains [88], however, the difference was not as marked. As the assembled Benchimol genome is only 65Mb in length, especially when it was predicted to be 161Mb, it is highly possible that many genes and repetitive regions have been missed and, therefore, that the Dabrowska *et al.* SNP data are incomplete. The aim of this chapter was to provide a complete genome sequence and improved assembly compared to the existing genomes, using long read sequencing technology.

2.1.4 The *T. vaginalis* Genome

As *T. vaginalis* is the closest, better relative of *T. foetus*, it is sensible to examine the similarities between their genomes. Furthermore, comparisons between the two would help to evaluate the reliability of the *T. foetus* genome and annotation. In 2007, the *T. vaginalis* genome sequence was sequenced [125]. Due to its size (estimated 160Mb) and high percentage of repetitive regions (approximately 65%) the sequence was highly fragmented, containing 17,290 scaffolds and an N50 of 68,338. This fragmentation and the number of transposable elements present has made key features of the genome and its true size difficult to determine, as the repetitive nature of some regions could lead to these being over expanded especially when short DNA sequence reads are used. It is thought that the high number of repeats is due to a genetic bottleneck in *T. vaginalis* history [125]. Sixty-five genes were determined to possess introns. All eukaryotes possess introns [245] [246] and there can be large ranges of intron density between species. This appears low compared to many eukaryotes, such as in humans where it has been estimated there are on average eight introns per gene [245]. However, this comparatively low number is not unheard of in other, more primitive branches of eukaryotes. *Giardia*, for example, is predicted to have five spliced genes [247] and only 6% of *Saccharomyces cerevisiae* genes are thought to contain introns [247], however, this still equates to over 2000 genes. Of the protists studied by McGuire *et al.* (2008) [247], including *P. falciparum*, *E. histolytica* and *P. tetraurelia*, all possess over 2,800 spliced genes in the genome, apart from *E. histolytica* which had 599.

The estimated genome size of 160Mb also differs from the actual reported genome assembly of 176,441,227 bases (176Mb). A core set of 60,000 genes was identified, however, this includes novel transposable elements and repeat regions and the number of evidence-supported genes is 25,949 [125]. The transcription machinery appears to be more metazoan than protistan, this could be due to lateral gene transfer (LGT) between eukaryotes and prokaryotes [20]. It is thought that many of these genes have been obtained from Bacteroidetes-related bacteria [125] which are found in intestinal flora. The cell membrane is thought to be covered with lipophosphoglycan (LPG) and

this may be associated with adherence [67]. There are also several potential LGTs that are involved in sialic-acid synthesis, which has been reported to be present on the *T. vaginalis* cell surface.

Over 650 BspA proteins were identified in the genome [125] which are known to be cell surface expressed in bacteria and can mediate cell adherence. Cytolytic effector proteins have also been putatively identified upon adherence of *T. foetus* with host cells [248]. It is thought that this may be in order for the parasite to acquire nutrients such as iron from the host [123].

2.1.5 *T. vaginalis* Cell-surface Antigens

The creation and publication of the *T. vaginalis* genome [125] sequence has proved invaluable to identifying putative cell surface antigens [249]. Due in part to this genome sequence, several cell-surface adhesin proteins have been identified. This has allowed for many studies into the roles and regulation of these proteins [249].

The *T. vaginalis* lipophosphoglycan (TvLPG) is a known virulence factor [20] [250]. These cell-surface adhesins are a key area being examined for *T. vaginalis* vaccine candidates. When pregnant women, both *T. vaginalis* positive and *T. vaginalis* negative, were subjected to TvLPG monoclonal antibodies, produced to conserved regions of the TvLPG [250], there was an immune response in the infected women. Additionally, all eight strains of *T. vaginalis* that were tested produced a response, showing that the immunogenic epitopes are conserved within the populations.

Cadherine-like cell surface proteins (CLP) have also been identified [251] and were found to be upregulated upon interaction with host cells. One particular CLP was found to be upregulated 2.5 fold in strains of *T. vaginalis* with high attachment rates and high cytotoxicity [251] and so it is thought to have major roles in parasite-cell interactions. This CLP is of particular interest as its induction appears to be mediated by the presence of environmental calcium, rather than iron or any other commonly known stimulus [252].

2.1.6 Other Trichomonad Genomes

Although *T. vaginalis* is the most well-studied trichomonad, due to its role in human disease, there are several other trichomonad species that are more closely related to *T. foetus*. These include: *T. mobilensis* [253] and *T. augusta* [254]. Unfortunately, genomes have not yet been assembled and annotated for these or any other trichomonads apart from *T. gallinae*.

Using Illumina sequencing, the *T. gallinae* genome has recently been sequenced [255]. It was found to have a genome size of 54,799,485 (54Mb) in 11,704 contigs with an N50 of 20,741 and a GC content of 33.77%. 27,000 genes were found and annotated. This is much smaller than *T. foetus* and has a much lower gene count.

2.1.7 Transcriptomes

A transcriptome consists of cDNA sequences for all gene transcripts present at a particular time in a cell, tissue or organism. Thus, a transcriptome provides a quantitative account of gene expression. RNA sequencing ('RNA-Seq') results are often closely correlated with metabolite results, thus giving more information about pathways [256] [257]. Transcriptomes are also essential in comprehensive genome annotation, since mapping of RNA-Seq reads onto the genome reveals the position of genes that are small or heavily spliced, which might otherwise be difficult to find. However, it is estimated that only 5% of the genes in a genome are transcribed at any particular time [258]. Therefore, to obtain an accurate gene annotation or estimate of gene expression a wide range of cellular conditions need to be assayed. In this chapter, multiple transcriptomes for *T. foetus* are presented, each representing gene expression in different environmental conditions, and providing as close to a full repertoire of expressed genes as possible.

Transcriptomes are being increasingly used as a factor in antigen selection and preliminary vaccine design [259] [260]. For example, the use of transcriptomes, along with other 'omics' methods has been used by Tuju *et al.* [261] (2017) in a reverse vaccinology attempt for a malaria vaccine. This allowed quantification of stage specific mRNAs to be identified as assisted with reducing the number of candidates. Zawawi *et al.* [260] (2020) used the expression of transcripts to decide which genes to test for MHC-II binding epitopes when identifying vaccine candidates for whipworm infections [260]. Transcriptomes have also been used in COVID-19 work, to identify potential T-cell and B-cell epitopes in the SARS-CoV-2 genome [259]. In this chapter, transcriptomes produced under a variety of environmental conditions will be used to identify genes that are constitutively expressed in both *T. foetus* life stages and to identify and confirm the placement of ORFs in the genome annotation.

2.1.8 The Pseudocyst

The *T. foetus* pseudocyst is suggested to be involved in virulence or as an infective stage during bovine trichomoniasis, and similar cell types are known in other trichomonad species. As stated in

the general introduction (Section 1.1.3), *T. foetus* cells can exist in either trophozoite form (pear shaped) or in a pseudocystic/endoflagellar form [9] [29] [262]. It would appear that the difference in morphology occurs depending on the environmental conditions, with cells in favourable conditions displaying trophozoite form and those under less favourable or stressful conditions displaying the pseudocyst form (Figure 1.4). Given the difference in morphology it would seem likely that a corresponding change in gene expression will also occur. This development change is explored through transcriptomic analysis in this chapter. Cells can be induced to form pseudocysts in various ways; by decreasing the environmental temperature [29]; the depletion of iron [263] or the addition of drugs that affect the microtubules of the cell, such as colchicine or nocodazole [262]. Additionally, it was seen that pseudocyst induction due to the low temperature was reversible, unlike that of the microtubule affecting drugs. However, there may be many more environmental stress conditions by which pseudocysts are produced that are currently unknown.

A vaccine for *T. foetus* would need to be able to target both life stages, particularly if one stage is key in the pathogenicity of the organism. It is important to try to induce pseudocysts and stress responses in *T. foetus* in a variety of environmental conditions if possible as certain conditions may induce specific responses. In this way it will be possible to see a wide range of potential genes in the genome.

2.1.9 Sequence Networks

Networks are used to look at sequence similarity, interactions and relationships between genes. They are a visual way to summarise very large data-sets and can illustrate and highlight key relationships and sequence superfamilies [264]. Unlike phylogenetic trees they do not require very large multiple sequence alignments. Sequence networks can be used to look at diversity between species, for example, in the case of identifying evolutionary relationships between different organisms.

Networks can also show putative homologues and can include non-homologous sequences in analysis [265]. They can also show trends related to function [264] and can show high levels of similarity in primary sequence which can identify core genes across species [266] even when secondary or tertiary structures appear different. Phylogenetic structure can be implied or inferred from a sequence similarity network and can be used to assist with creating phylogenies [264] [266] [267]. Additionally, the functional data can be used to form or guide experimental hypotheses about the relationships between genes and gene families. Atkinson (2009) [264] used three groups of proteins with different structural and functional diversity: G-protein couple receptors (GPCRs), kinases and

the crotonase family of enzymes, and combined their sequence information with BLAST identity scores. This information was used to produce networks which showed clear clusters of related genes and were also in agreement with the known structures and functions of the gene families. It was also possible to identify very distant relationships between members of the GPCR group. The networks become more accurate and reliable as the number of entries increases because the connection between those entries increases. [266][265]. It is also easy to recompute the network if parameters are changed, for example, BLAST e-values or length of queries.

It is the aim of this project to use these networks to visualise a set of *in silico* cell surface protein predictions and to also infer function by using both *T. foetus* and *T. vaginalis* genomes. This is in the hope of identifying a function in the case of genes that currently have no known functions or homologs and to identify even distantly related genes within and between the two species. These networks will be able to show gene clusters and gene families that will be used for later steps of the reverse vaccinology pathway.

2.1.10 Aims and Objectives

The aim of this chapter is to assemble and annotate the *T. foetus* genome from PacBio sequence reads and produce an *in silico* cell surface proteome comprising all genes predicted to encode cell surface proteins. Reducing the whole genome set to a small cohort that can be validated proteomically is the first step in the reverse vaccinology process.

1. Using various computer programmes, the genome will be assembled several times using different parameters to create an optimised assembly.
2. The ORFs will be annotated by computational means and checked by manual curation to create a gene set.
3. *T. foetus* transcriptomes, produced under a variety of environmental conditions will be further used to identify and confirm ORFs.
4. An *in silico* cell surface proteome will be created by identifying all putative proteins containing a predicted N-terminal signal peptide and a predicted C-terminal transmembrane domain.

2.2 Methods

2.2.1 Culture Maintenance

The Belfast strain of *T. foetus* was grown in Diamond media [268]. The cells were kept at 37 °C and were passaged during log phase (24-48 hours after passage) when they were at an approximate density of $5x10^6$ per ml . Unless stated, otherwise all experiments and analyses were performed on the Belfast strain of *T. foetus* ATCC strain 3016. The parasites were grown in 15ml tubes, 1 litre of Diamond media contained: 15g tryptone, 12g yeast extract, 5.5g glucose, 2.5g sodium chloride, 0.5g L-cystine, 0.5g sodium thioglycollate, 1% penicillin/streptomycin (1,000,000 U), 100ml horse serum and 0.75g agar.

2.2.2 Assembly and Size Estimation of the *Tritrichomonas foetus* Genome

DNA sequence reads were produced at the Centre for Genomic Research (University of Liverpool) using the RSII machine; eight SMRT cells produced 249,576 10kb reads [224] [269] [270]. The data were assembled *de novo* with SMRTportal using Hierarchicial Genome Assembly Process (HGAP) 2 which is optimised for accuracy. The parameters were changed to see how that affected the predicted size of the genome so a final value could be established (Table 2.1).

The changed parameters included: max divergence, predicted size and minimum seed read length. Assemblies were compared according to the total size, number of contigs and N50 value that they returned. N50 statistic defines assembly quality in terms of contiguity. The N50 is defined as the minimum contig length needed to cover 50% of the genome. It means half of the genome sequence is in contigs larger than or equal the N50 contig size. The size estimated by Jellyfish was specified when assembling the genome using the Hierarchical Genome Assembly Process (HGAP) [269] within the SMRT Portal server. The genome was assembled several times using the SMRT Portal server to reach a consensus (ENA study PRJEB41567).

Parameter	Value
Expected genome size (MB)	100
Minimum Seed Read Length	8,000
Maximum Divergence (Total)	10

Table 2.1: Parameters used in final *T. foetus* genome assembly with SMRT Portal

The genome was annotated using Artemis with the *ab initio* predictions from SNAP [232] and BRAKER [233]. The transcriptomic data were used 'train' the BRAKER annotation and to predict both genes and splice sites when manually annotating in Artemis. Jellyfish used pre-assembled reads from the SMRTportal values. Kmer values of: 17,19 and 21 were used in the genome size prediction. The expected genome size predicted by Jellyfish were then used as the estimated sizes in the SMRTportal assembler and the assembly was rerun.

SMRT Portal

HGAP process

The reads produced by PacBio are filtered for a minimum length and quality and the adaptors are removed. The reads are mapped, short to long to produce the preassembly before being assembled by the celera assembler OLC to form the draft assembly. Finally, the assembly is 'polished' meaning the reads are mapped to a current assembly and errors can be corrected using quiver. HGAP2 is optimized for quality rather than speed (as in the case of HGAP3).

Canu

Canu is a successor of the celera assembler [271]. Canu, a hierarchical method, uses multiple rounds of overlapping and alignment before assembly to improve the reads. This is particularly important for newer sequencers where longer reads are produced with a higher error rate.

2.2.3 Identification of ORFs

In order to find the open reading frames in the assembled genome several different programme were used. One contig was annotated manually to show both ORFs and splice sites and used as a 'training set' for all further computational programs. In the first instance the program 'AUGUSTUS' [231] was used, however, it was not possible to get the program to work effectively.

An attempt was made to use the companion server [272] to annotate the genome. Companion uses a reference-based approach to identify genes. It involves orientating the contigs and using Rapid Annotation Transfer Tool (RATT) [273] to transfer highly conserved sequences from the reference to the new genome. AUGUSTUS [231] is also used within the server to annotate genes *de novo* using the reference genome as a training set. Unfortunately there was no reference genome close enough to *T. foetus* to use. The *T. vaginalis* genome was trialled, however, the genome did not have a high enough similarity to that of *T. foetus* and there are many trichomonads more similar but

without annotated and sequenced genomes. The Gene Locator and Interpolated Markov ModelER (GLIMMER) online server [274] [275] was also used and 166,143 ORFs were found in the genome. The GLIMMER server works by using interpolated Markov models (IMMs) to distinguish coding regions from non-coding regions. However, GLIMMER did not identify splice sites so the high number of ORFs compared to the manually curated number (see below) could be due to individual exons being counted as genes.

The programs used for the final computational annotation were BRAKER and SNAP. One large contig was annotated manually to create a training set. Additionally, BRAKER also used mapped *T. foetus* transcriptomes to identify the splice sites. This had the advantage of showing where many of the splice sites started and ended, whereas SNAP only showed the individual ORFs. The genome was then annotated in Artemis [276]. The BRAKER and SNAP annotations were used as a guide to find the reading frames which were then manually curated to check their placement.

2.2.4 Manual Curation

After the computational methods were used to identify the ORFS, they had to be curated by hand in Artemis to check the start and stop codons and splice sites were in the correct place. This was to identify correct genes due to the discrepancies between the different computational programs. In order to identify accurate ORFs; the GC content, frame usage and transcriptomic data were used. Clear ORFs were marked as were splice sites, using the consensus sequences: 3' consensus sequence: CAG—G and 5' consensus sequence MAG—GTRAGT where M is A or C and R is A or G [277]. This is the standard splice consensus sequence, it is possible the *T. foetus* uses non-standard splicing sequences.

2.2.5 Gene Annotation

Initially, ORF sequences were examined in Blast2Go, using BLASTx to compare each against the NCBI nr database [278]. This showed which genes had homology with existing, named proteins, which were conserved hypothetical, ie. they were homologous to existing, unnamed proteins, and which were unique hypothetical, i.e. they had no homology with other genomes. The predicted protein sequences of each ORF were also examined using several other programs for structural features, including InterProScan, KEGG, SignalP and TMHM

Blast2GO

Blast2GO is a suite of bioinformatics tools within an easy to use graphical user interface [279] [280]. Blast2Go can assign GO terms to sequences even if they are unknown. It is also possible to use a range of bioinformatic tools such as BLAST and InterProScan within the BLAST2GO software. It can also produce statistics and and aid in visualisation of results.

InterProScan

InterPro [281] [282] provides information about protein motifs and signatures. It uses a variety of databases to do this including: PANTHER, prosite and pfam [283] [284] [285]. The use of many different databases allows a comprehensive result to be determined as each database has different strengths and criteria and it is due to this that helpful protein annotations can be derived.

KEGG

KEGG [286] [287] allows for the high level functions of genes to be elucidated and the overall biological systems present in a given organism. It also allows for interpretation of biological function of genes in a genome and can infer biological pathways. It comprises several categories including glycans, drugs and diseases [288].

TMHMM

TMHMM [289] [290] predicts transmembrane helices based on a hidden Markov model (HMM), it has been shown to be able to correctly predict transmembrane helices (TMH) 97-98% of the time [290]. It is also able to distinguish between soluble proteins and membrane proteins 90% of the time. It identifies these TMHs by combining hydrophobicity and charge bias analyses [291].

SignalP

SignalP [292] predicts signal peptides. Signal peptides target proteins for translocation across the cell membrane using an artificial neural network approach [291]. It identifies hydrophobic-α-helical regions which are a strong indication of signal peptides [291]. Knowing whether a protein has a signal peptide can also help identify the orientation of a transmembrane helix; if an SP is present, the N terminus of the TMH has to be on the non-cytoplasmic side.

PredGPI

PredGPI [293] predicts GPI-anchors. GPI-anchors are a form of lipid anchor for cell surface proteins [294] and GPI-anchored proteins can form a major type of cell surface protein in eukaryotes, particularly protozoa. PredGPI uses Hidden Markov Models and support vector machine learning to predict the presence of the anchor and the ω-site, the site of cleavage of the terminal C-terminal residue- a key post-translational modification in GPI-anchor formation.

Orthofinder

Orthofinder can be used to find homologues between species and can identify putative gene families [222], forming orthogroups. Identifying this homology can allow us to infer function or classification of unknown genes by their similarities to known ones, such as between *T. foetus* and *T. vaginalis*.

2.2.6 Initial Transcriptome Mapping with Trinity

The initial Trinity assembly was performed to provide a check that the transcripts were mapping correctly and to identify the top BLAST hits. The assembly was run through Blast2Go to see what the species distribution was. The top two hits were *T. foetus* and *T. vaginalis*. As these were the top species hits it was reasonable to assume the data was from the expected organism and mapping correctly. The Illumina reads can be used to assemble both DNA and RNA sequence data.

2.2.7 Environmental Conditions for Transcriptome Creation

After automated genome annotation and manual curation of gene models, many putative ORFs lacked evidence for transcription, that is, RNA-Seq reads provided zero coverage of those loci. This could be due to the fact that certain genes are only expressed under certain conditions, other than those encountered in normal cell culture. In order to maximise the representation of genes in RNA-Seq data, a range of different environmental conditions were applied to cell culture: high and low temperature; high and low pH; oxidative stress and nutrient depletion (ENA study PRJEB39462).

pH

The pH of the environment is likely to invoke a stress response. In cattle, the pH of the vagina ranges from 7.3-7.6 with a median of 7.5 [295]. Therefore, it would seem logical that the optimum pH for the growth of *T. foetus* would be around this value. If the pH was increased or decreased

from this value then growth could be affected and the cells could become stressed and could induce pseudocyst formation or induce different genes. There are known changes in pH of the cow vagina depending on pregnancy status [132]. Additionally, the gastrointestinal tract, which *T. foetus* can colonise in cats, can also have fluctuations in pH, the severity of which depending on the section of the tract. Therefore, *T. foetus* must have adaptations to these changes in conditions.
The pH values of 6 and 8 were chosen as this is outside the average normal range of the growth environment for *T. foetus* but is not so extreme as to kill the cells and prevent any growth.
The normal pH for prepared Diamond media is 7.3. The standard Diamond media was used and the pH altered to the required value. One tube of 15ml was split into three, with 5ml in each sample containing approximately $6x10^7$ cells per ml. One sample was used for high pH, one for low pH and one as a control. The pH was measure using vw pH strips and the pH was altered by the addition of NaOH or HCl. Samples of 5μl were taken at time intervals and the motility, morphology and mortality of the cells was noted.The resulting pHs of the media were: high pH-8.5-9, low pH-5.5-6, control pH-7.3. This was repeated three times and RNA was extracted and sequenced for each sample.

Low temperature

In order to induce pseudocysts, 15ml tubes of cells were placed at 4°C for 4 hours and the proportion of trophozite and pseudocyst cells were counted every 30 minutes for 2 hours and then every hour for the following 5 hours. Samples were taken when the proportion of pseudocysts was over 90% and RNA was extracted and sequenced (Figure 2.8).

High temperature

Cells grown in Diamond media were grown to log phase and split into 3 samples of 5ml. The temperatures chosen were: 35°C, 42°C and 46°C. Cells were incubated at these temperatures and samples were taken every 15 minutes for the first 30 minutes and then every 30 minutes for the next 5 hours. The morphology, motility and mortality of the cells was observed over this time (Figures 2.13 and 2.12).

Oxidative stress

It is known that an increase in oxygen can affect the levels of ethanol, CO_2 and H_2 produced by the cells [296] and can also affect the growth rate of the cells themselves [297]. There are also

bacteria, commonly found in the urogentital tract, such as *Lactobacillus*, that can produce H2O2 [298]. Cows have lower levels of *lactobacilli* in their urogenital tract than humans, however, there are other bacteria that produce lactic acid and compete for resources, such as the closely related genus *Pediococcus* [298]. In order to induce oxidative stress cells were grown to mid-log phase and hydrogen peroxide was added.

Initially stock solutions of hydrogen peroxide, 6mM and 10mM were made up in diamond media from 30% hydrogen peroxide solution.

The H_20_2 was added to $6x10^6$ cells to give final concentrations of 300μM and 500μM. The cells were then grown and samples taken every 30 minutes (Figure 2.9).

Starvation stress

The gastrointestinal or urogenital tracts inhabited by *T. foetus* contain a wide range of microbial flora and fauna. The presence of other microorganisms will cause competition for nutrients and *T. foetus* will have to either outcompete other organisms or adapt to depleted nutrients, such as glucose. In order to induce different stress responses due to the depletion or absence of nutrients, three different types of Diamond media were made up- each lacking a particular component from the original recipe:

1-Diamond media, without 10% supplemented horse serum

2-Diamond media, without tryptone and L-cystine

3-Dimond media, without glucose

Cells were visualised 24 and 48 hours post media change and a trypan blue assay performed to calculated percentage cell death. The motility and morphology was also examined.

Colchicine

Cells were grown at 35°C and the proportion of trophozoite and pseudocyst cells were counted every hour for 7 hours after the initial addition of 1.5mM colchicine (Figure 2.8).

2.2.8 RNA Extraction and Sequencing

Cells were harvested at a density of $1x10^7$ cells by centrifugation at 800 rpm for 5 minutes. The cells were washed in PBS (Thermofisher) and centrifuged again at the same speed for 5 minutes. RNA was extracted using the Qiagen RNeasy kits as per the manufacturer's instructions. To quantify the amount and concentration of RNA produced, the samples were run out on a 1% agarose gel and then

analysed with Qubit and nanodrop instruments noting the 260/230 and 260/280 ratios respectively. The libraries were prepared either by the CGR using poly-A selection or using Serapure beads and the NEBNext Ultra prep kit.

Serapure beads were prepared using a protocol derived from Rohland (2012) [299]. Libraries were created using: NEBNext Ultra II Directional RNA Library Prep kit, NEBNext Poly(A) mRNA Magnetic Isolation Module and NEBNex multiplex oligos for Illumina. The libraries were sequenced at the CGR using the Illumina HiSeq and Novaseq platforms.

2.2.9 RNA-Seq Analysis Pipeline

Raw RNA reads were trimmed for the presence of Illumina adapters using Cutadapt v1.2.1 [300] the 3' end of any reads that matched the Illumina adapter for 3bp or more was trimmed. The reads were trimmed using Sickle v1.200 [301] using a minimum window quality score of 20. Reads that were less than 15bp after trimming were removed. The resulting fastq files were uploaded to the Galaxy server [302] and mapped to the completed genome assembly sequence using the program Hisat2 [303] [304].

The bam file produced by Hisat2 was sorted and indexed using samtools [305] and used in conjunction with Artemis to visualise RNA transcripts across the genome. The initial Hisat2 mapping output bamfile was also run through FeatureCounts[306], again on the Galaxy server, to produce counts tables. The counts tables were then used as input files in R Studio [307] [308] and the differential expression analysed using the DeSeq2 package [309]. The Hisat2 program also produced output files stating how much of the sequence was mapped to the genome (Table A1).

2.2.10 *In silico* Cell Surface Proteome

Cell surface localisation is the first and principle criterion for plausible vaccine antigens in this thesis. Therefore, a long-list of vaccine candidate antigens was obtained from the *T. foetus* gene set by selecting all genes with a predicted cell-surface protein. These comprised amino acid sequences with predicted N-terminal signal peptides (SP), either alone (i.e. putative secreted) or in combination with one transmembrane domain (TMD) (i.e. type-1 transmembrane protein), or multiple TMDs (i.e. multi-spanning protein) or GPI anchor (i.e. membrane tethered). All of these proteins are assumed to be located on the plasma membrane because of the SP.

Given the terminal position of SP, TMD and GPI motifs within gene models, and allowing for the

possibility that some gene models may be incomplete, each predicted cell surface gene sequence was compared to all *T. foetus* genes using BLASTP to identify any homologs that might be missing the key structural features in the annotation. Each gene was also compared to a database of all *T. vaginalis* genes using PSI-BLAST to identify homologs in the related parasite. PSI-BLAST was used as it accounts for protein secondary structure in its comparison, generally making comparison more sensitive, and thus, more likely to uncover weak similarity between species.

Having obtained a set of all putative *T. foetus* cell surface proteins and their *T. vaginalis* homologs, pairwise PSI-BLAST bit scores were calculated by comparing each entry to a BLAST database of all proteins. Self-matches were removed, as were , all those with a total amino acid sequence similarity of less than 30%, an e-value more than $1e^{-20}$ and where the length of the match was less than 150 amino acids. All other PSI-BLAST bit scores that exceeded these thresholds were extracted and used to create a sequence network.

Two files were created for use in Cytoscape [310], the input file and attributes file. The input file contains all sequence names (i.e. nodes in the network) and their pairwise PSI-BLAST bit score in comparison to other sequences, where a significant match was obtained (i.e. the edge weights in the network). The bit score was used because it is derived from the PSI-BLAST alignment score and is normalised relative to the scoring system used. This means the bit scores can be compared if different BLAST analyses are run. The attributes file contains information about the identity, origin and structure of each sequence. Nodes in the network were colour-coded by species and by the presence or absence of SP, TMD and GPI.

The network was manually curated to remove obvious errors or instances irrelevant to the analysis. For example, clusters derived from transposable elements or clusters in which only a single representative of many qualified as cell surface expressed by the criteria given above. In the latter situation, alignments were created to determine whether the SP was likely to be correct or was more likely to be due to an error in the original genome annotation. Genes were also examined to for interpro domains that made their annotation consistent with cell surface expression.

2.3 Results

2.3.1 Genome Assembly

The *T. foetus* (Strain Belfast) was sequenced on the PacBio RSII machine and initially assembled using SMRT portal, the proprietary PacBio assembly software, using Hierarchical Genome Assembly Process 2 (HGAP2) [269] under default conditions. This initial analysis provided a benchmark of assembly statistics such as contig size and number, and N50. Different parameters were then changed to see how they affected the overall size of the genome assembly (Figures 2.1, 2.2 and 2.3). The parameters that were changed were: expected genome size, minimum seed read length and maximum divergence.

The size of the genome annotation was expected to be around 150-200Mb, based on the prediction of 177Mb +/- 12 of Zubakova *et al.* [114] and the size of the *T. vaginalis* genome (160Mb) [15]. The predicted size using Allpaths-LG 2017 [227] was 161Mb for the *T. foetus* genome. The final assembly conditions selected were an expected genome size of 100Mb, a max divergence of 10 and a minimum seed read length of 8,000. The overall statistics for this optimised genome assembly were: 147Mb in size with an N50 of 82,179 and 2,776 contigs (ENA study PRJEB41567).

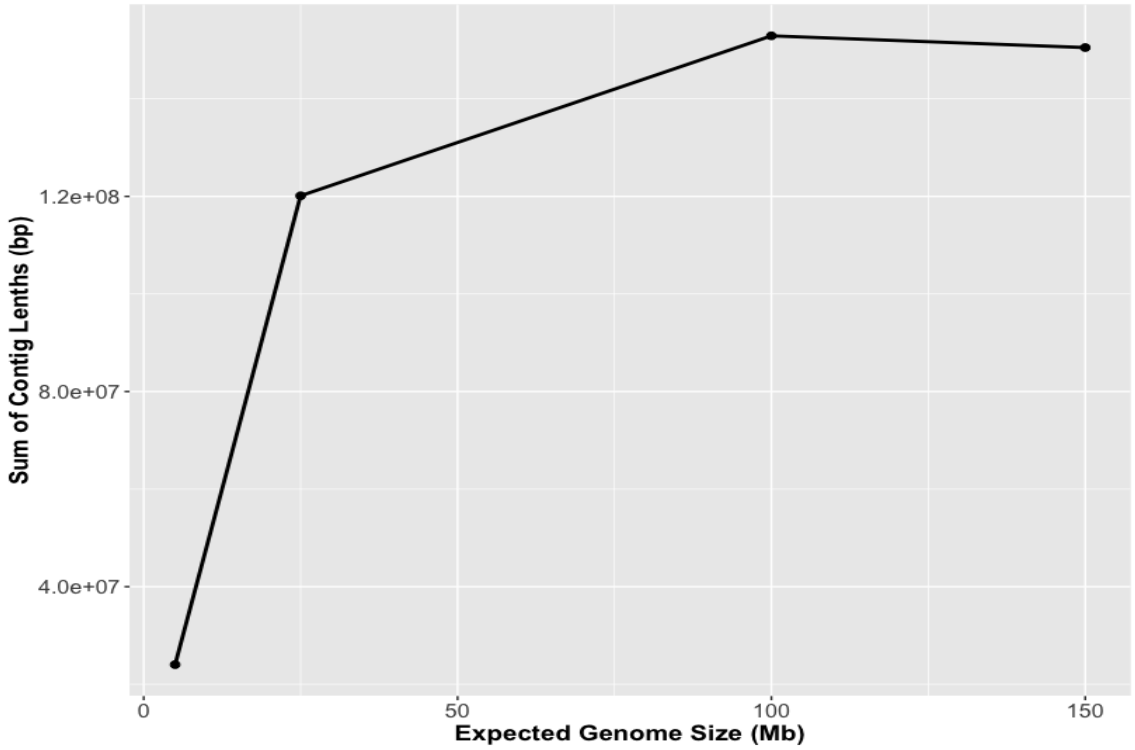

Figure 2.1: Identifying how the change in SMRT portal parameters affects the predicted genome size. This graph shows the effect that changing the expected genome size had on the sum of contig lengths where max divergence was kept at 30% throughout and minimum seed read length was kept at 8,000. All trials were performed using HGAP2 in the SMRT portal software.

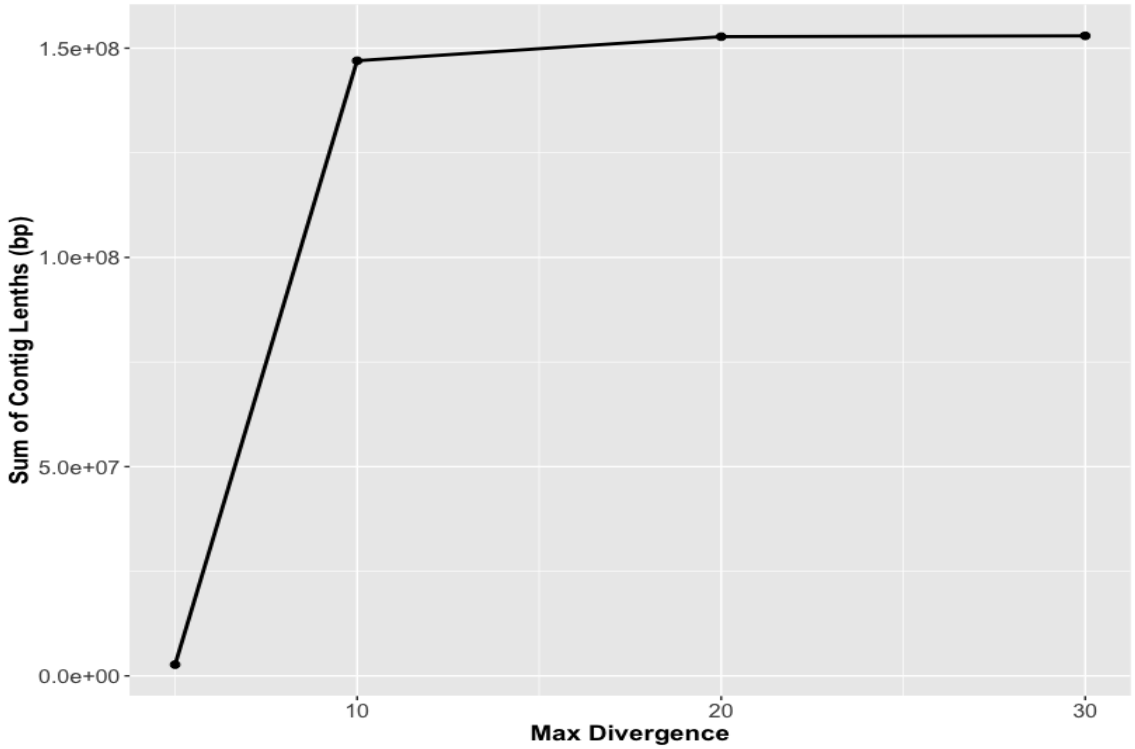

Figure 2.2: Identifying how the change in SMRT portal parameters affects the predicted genome size. This graph shows the effect that changing the maximum divergence had on the sum of contig lengths. Predicted genome size was kept at 100Mb and the minimum seed read length was kept at 8,000. All trials were performed using HGAP2 in the SMRT portal software.

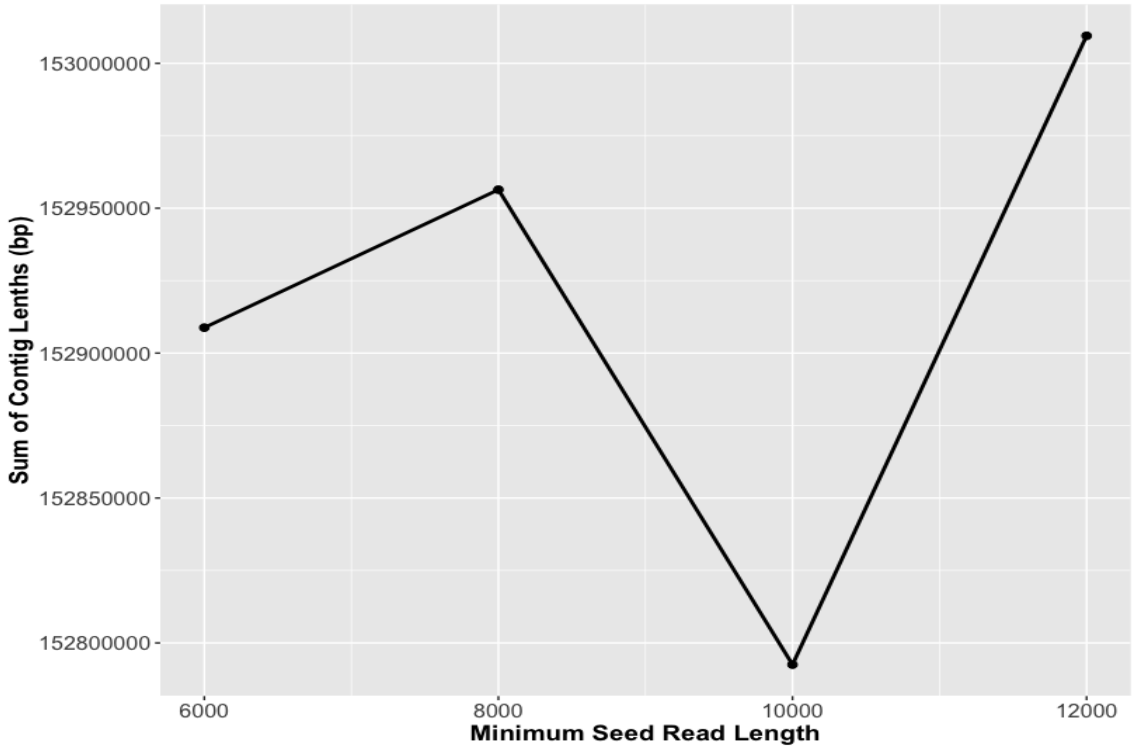

Figure 2.3: Identifying how the change in SMRT portal parameters affects the predicted genome size. This graph shows the effect that changing the minimum seed read length had on the sum of contig lengths. The predicted genome size was kept at 100Mb in all trials and the maximum divergence was 30% in each All trials were performed using HGAP2 in the SMRT portal software.

2.3.2 Transcriptome Assembly

An initial Trinity assembly was made of a *T. foetus* transcriptome that had been produced in September 2016. This was mapped onto the first attempt at genome assembly with SMRT portal to obtain a rudimentary genome annotation. All large open reading frames (over 200 amino acids) were selected. The genes were then compared to the NCBI nr database using BLAST2GO to see whether they appeared to be from the correct organism. The results are shown in Figure 2.4, which shows that top BLASTn hits come from *T. foetus* and *T. vaginalis* (nearly 50,000 and 25,000 hits respectively), providing further evidence that the genome is for the correct organism.

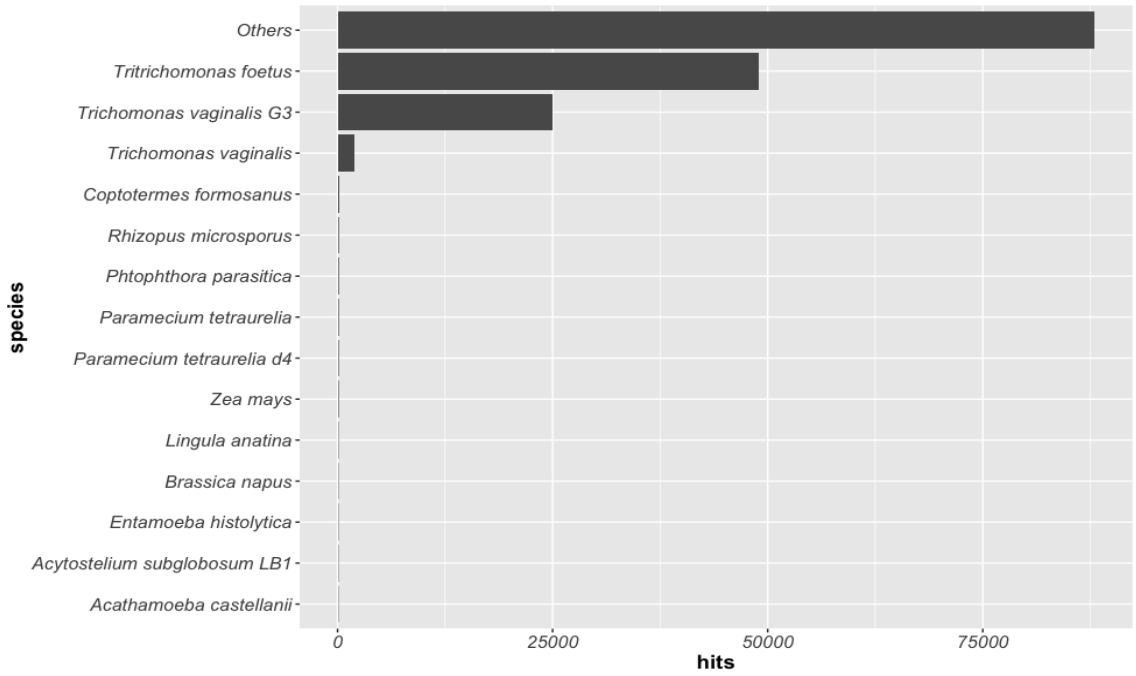

Figure 2.4: Species distribution of the Trinity assembly of the *T. foetus* genome. Data obtained includes the top 14 species (plus 'others') with BLASTn hits from the NCBI nr database with the highest sequence similarity to the *T. foetus* transcriptome sequences. The BLAST search was performed via Blast2Go pro.

2.3.3 Genome Annotation

The training set for automated annotation was created through manual annotation of ORFs and splice sites consensus sequences, as previously defined [277], in a single 64,4735bp contig; this produced 251 gene models. Two automated annotation programs that employ different approaches to ORF and splice site discovery were used: BRAKER [233] and SNAP [232]. Overall, SNAP identified 134,803 exons over the genome and BRAKER identified 111,040 exons and introns. BRAKER also predicts splice sites, which were manually added to the annotation. All the exons identified by the software were manually curated to check that they started and ended in the correct places and contained the correct number of exons. After manual curation, 84,725 genes had been found.

2.3.4 General Features of the Genome

When tRNAscan [311][312] was run using the genome assembly data, 251 tRNA sequences were found. 39 contigs were found to contain multiple tRNAs with the largest number in any one contig being 5. When RNAmmer [313] was run, a total of 101 sequences were found. Of these: 47 were

Attribute	Result
Size (bp)	147,002,103
Number of Contigs	2,776
GC%	32.74
Genic component (%)	58.8
Non-genic component (%)	41.2
Gene density (per kb)	0.576
Protein coding genes (Total)	84,725
Protein coding genes (Spliced)	17,602
Protein coding genes (Non-spliced)	67,123
Mean number of exons	1.26
tRNA genes	251
rRNA genes	101

Table 2.2: General features of *T. foetus* Genome

8S, 27 were 18S and 27 were 28S. When the number of *T. foetus* RNA genes were compared to *T. vaginalis*, *T. vaginalis* has a much higher number [314] with 668 rRNAs and 468 tRNAs. This could mean that some RNA encoding genes have been missed by the program or have been annotated incorrectly

In many cases, obvious ORFs in the genome sequence had no evidence of gene expression after mapping of RNA-Seq reads. This suggests that only a minority of genes are expressed at any one time, and some genes will only be expressed under specific conditions, not found in *in vitro* cell culture. In order to increase transcriptomic evidence for annotated gene models, and to discover more, cell cultures of both trophozoites and pseudocysts were established in various environmental conditions, RNA-Seq data from cells cultured in high and low temperature, high and low pH, oxidative stress and (latterly, see Chapter 4) the presence of a host cell monolayer. All of these transcriptomes were used during manual curation of predicted gene models to produce the final number of genes.

Initial sequence analysis of the gene models using BLASTx in BLAST2GO against a non-redundant database found that 12,114 were hypotheticals, homologous to known proteins in the *T. foetus* genome that had been produced. Whereas, 42,730 were unique hypotheticals with no homology in other genomes, suggesting that these could be *T. foetus* specific and 29,881 had BLAST hits to known proteins in other genomes .

GhostKOALA

GhostKOALA [288] [315] was used to search and assign KEGG pathways and functional annotations to the annotated genes of the genome (Figure 2.5). 9,803 proteins (11.6%) of the genome

mapped to 375 pathways. The pathways with very high numbers of proteins linked to them include: Lysine biosynthesis, acyl-coA biosynthesis and purine metabolism.

The majority of the defined taxonomy belongs to the parabasalids, which is the taxonomic group *T. foetus* belongs to (Figure 2.6). Only 18 KEGG modules are fully mapped to *T. foetus* genes and include pathways such as: glycolysis, the pentose phosphate pathway, glycogen degradation, fatty-acid elongation and pyruvate oxidation. Most of the completed pathways found were members of carbohydrate metabolism pathways. If modules that have all but one component mapped the number increases and includes: glycan biosynthesis, Co-enzyme-A biosynthesis and proline biosynthesis.

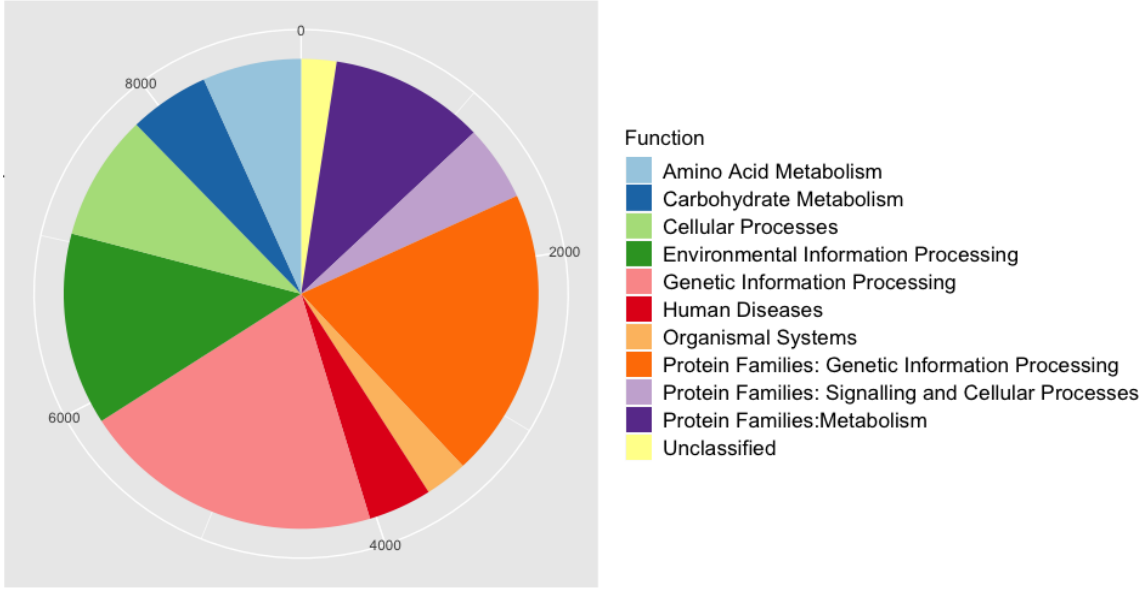

Figure 2.5: The functional categories and KEGG pathways identified by GhostKOALA searching of all annotated ORFs in the *T. foetus* genome. 9,803 ORFs out of 84,706 were successfully mapped to 375 pathways. These pathways were clustered into functional categories, the 11 with the most mapped ORFs are shown here.

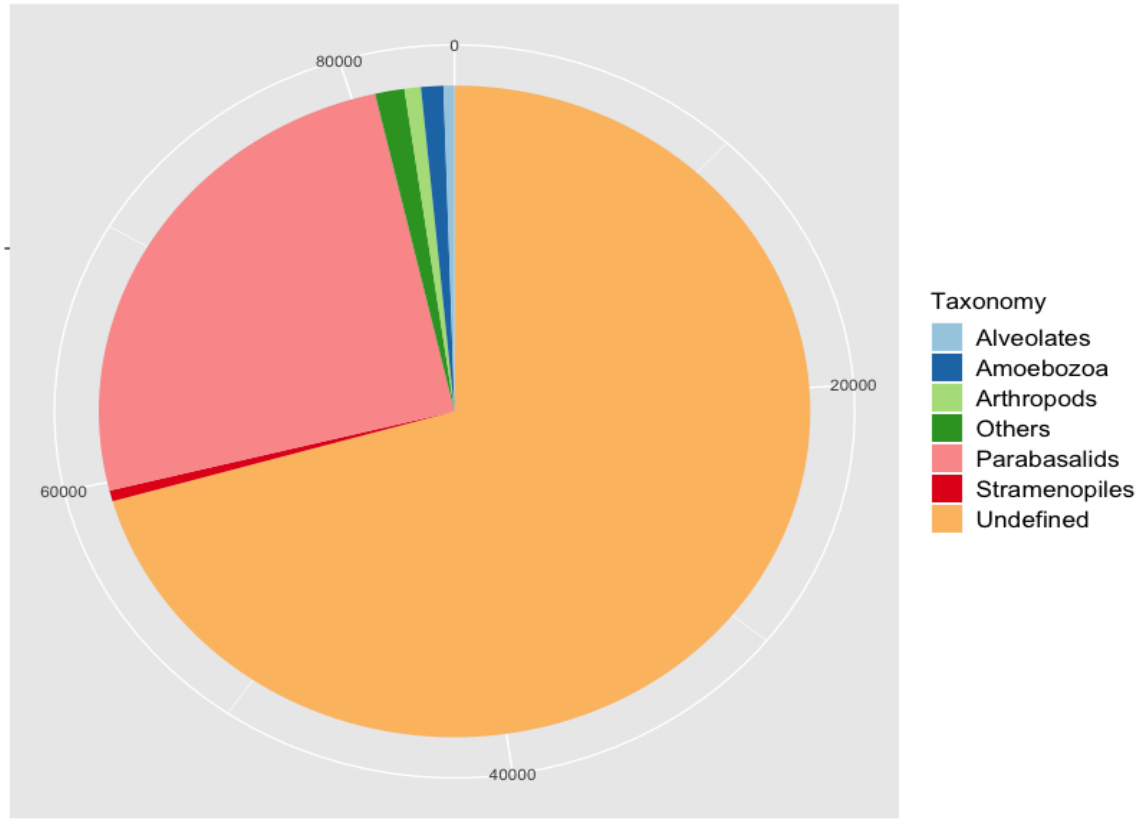

Figure 2.6: The Taxonomic results of a GhostKOALA search of all annotated open reading frames in the *T. foetus* genome. The majority of results were undefined with Parabasalids being the second most common after 'undefined'.

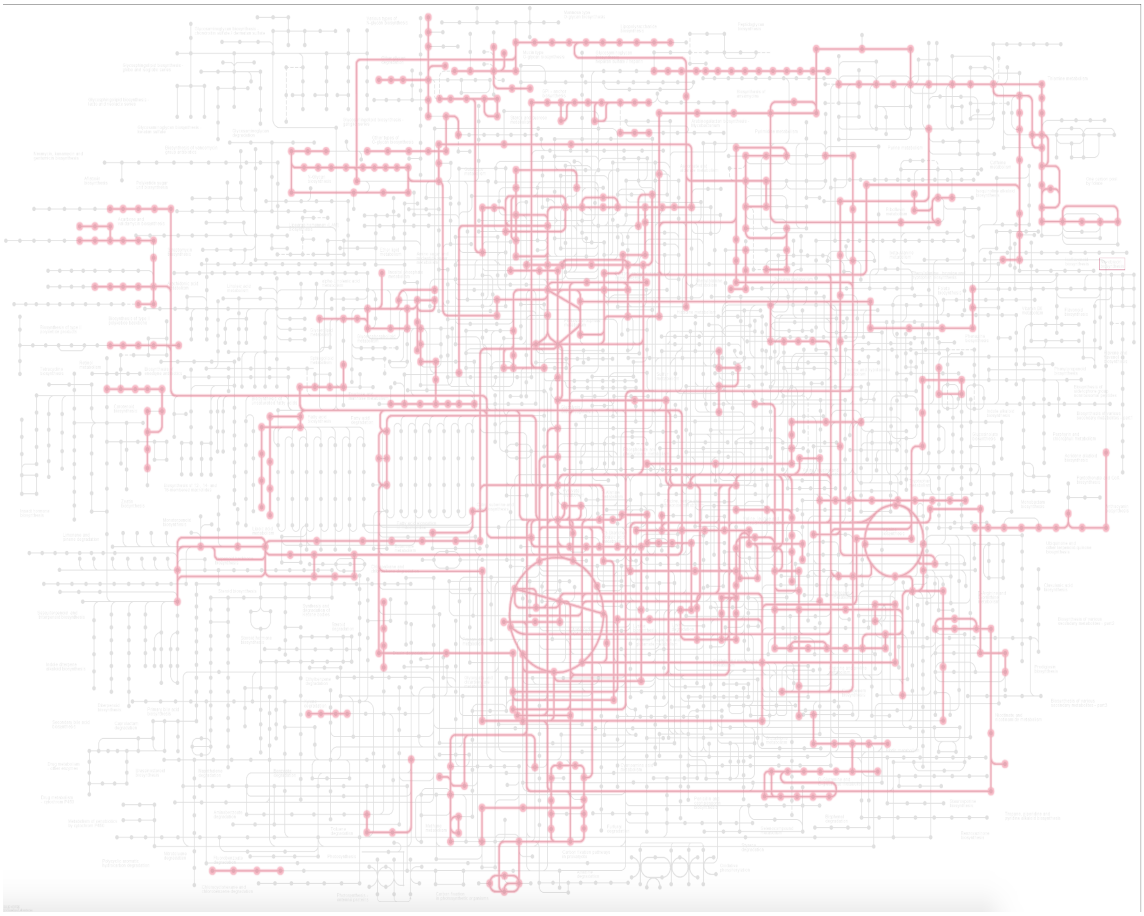

Figure 2.7: KEGG metabolism pathways. The mapped KEGG modules of all annotated ORFs in the *T. foetus* genome, highlighted in red, produced by GhostKOALA. Modules where at least one component is represented within the pathway have been highlighted.

2.3.5 Pseudocyst Induction

The proportion of pseudocysts formed over time under each of the environmental conditions, were calculated. The percentage cell death was also calculated as, in some cases, cell death occurred very shortly after the stress was applied. This was in order to identify the ideal time to extract the RNA so the results would show a clear response to the environmental stress without the cells dying and producing an inaccurate gene expression profile. The low temperature condition produced the most consistent pseudocyst response and a sample could contain over 75% pseudocysts after 4 hours and nearly 100% pseudocysts after 7 hours (Figure 2.8). Different environmental stresses induced different proportions of pseudocysts, so the RNA-Seq data was also used to examine if all pseudocyst samples induced the same gene expression profile or whether the induction method affected this.

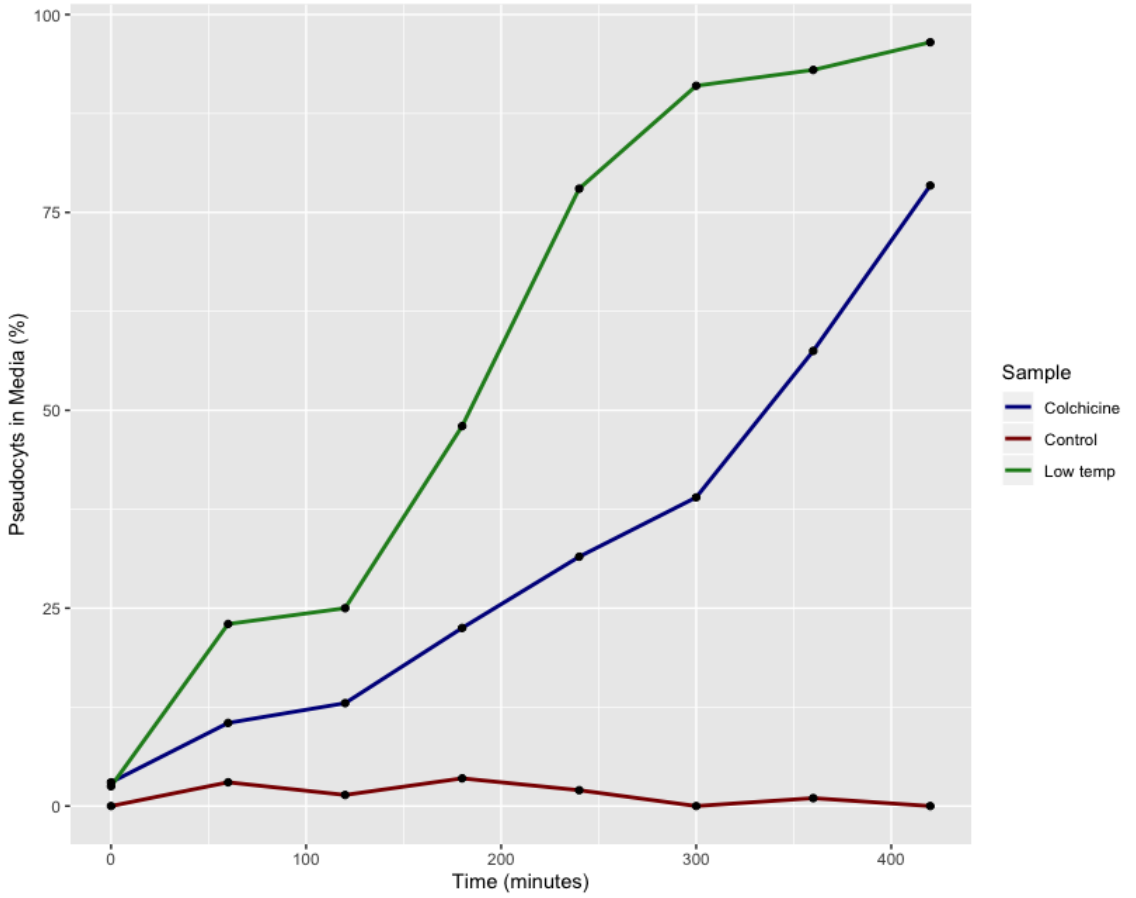

Figure 2.8: Proportion of *T. foetus* cells grown in Diamond media that are in pseudocyst form rather than trophozoite form when a low temperature stress (4°C) or an addition of colchicine was used. Samples were taken every hour and morphology was identified using light microscopy techniques [150] [16].

When oxidative stress was used to induce pseudocysts (Figure 2.9), the rate of pseudocyst induction increased when the concentration of hydrogen peroxide was increased from 300μM to 500μM. However, when a pH stress was applied, the proportions of pseudocysts formed was very low, never exceeding 7% of the total sample (Figure 2.10). Looking at the cell death proportions for the samples, this may be due to the fact that some cells were dying before they could form the pseudocysts. The death rate for the low pH samples (Figure 2.11) is much higher than the other samples, and the rate of pseudocysts also increases but only 5% higher than the controls. This also shows that the *T. foetus* cells can cope with a pH stress slightly higher than the norm more successfully than a low pH stress. When pH values that were below 5 and above 9 were used, the cells immediately died.

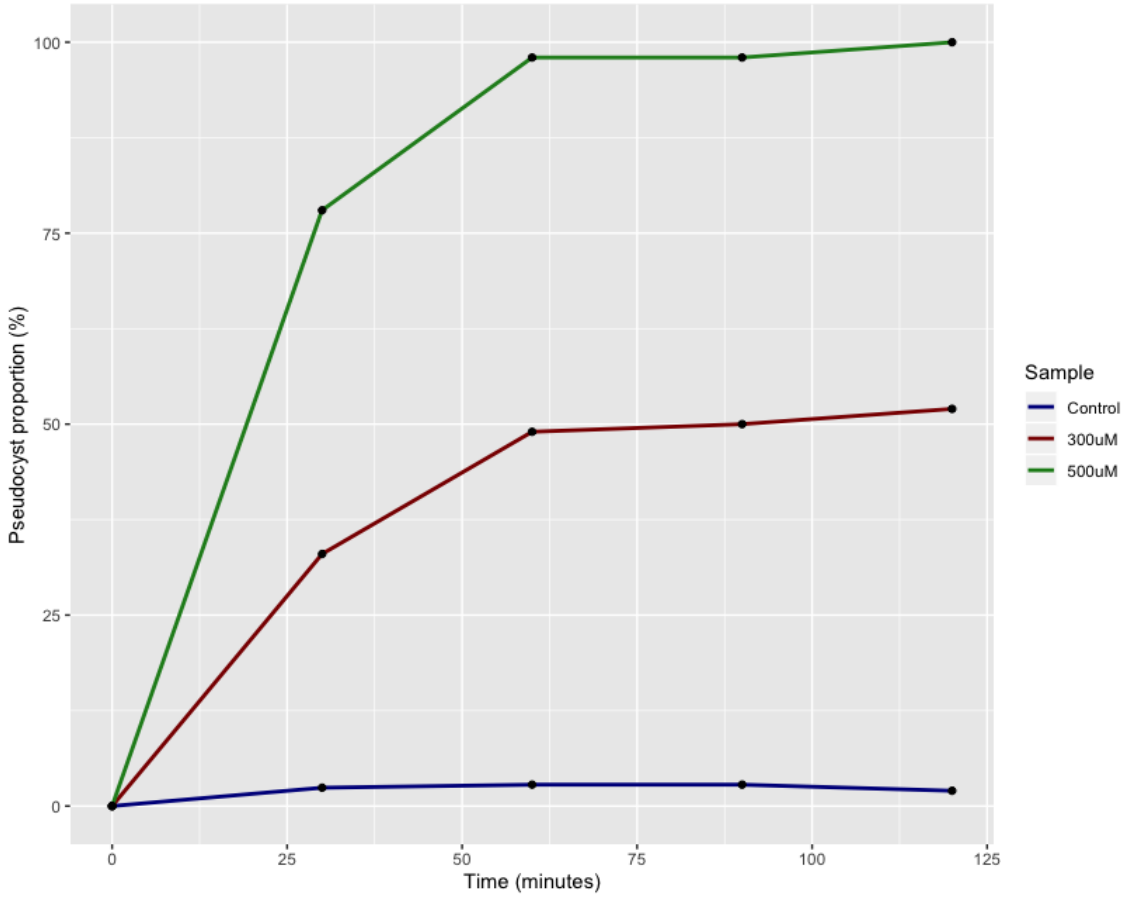

Figure 2.9: Proportion of *T. foetus* cells grown in Diamond media that are in pseudocyst form rather than trophozoite form when hydrogen peroxide is added to the media at concentrations of 300μM and 500μM compared to control cells where no hydrogen peroxide has been added. Samples were taken every 30 minutes. Morphology was identified using light microscopy techniques [150] [16].

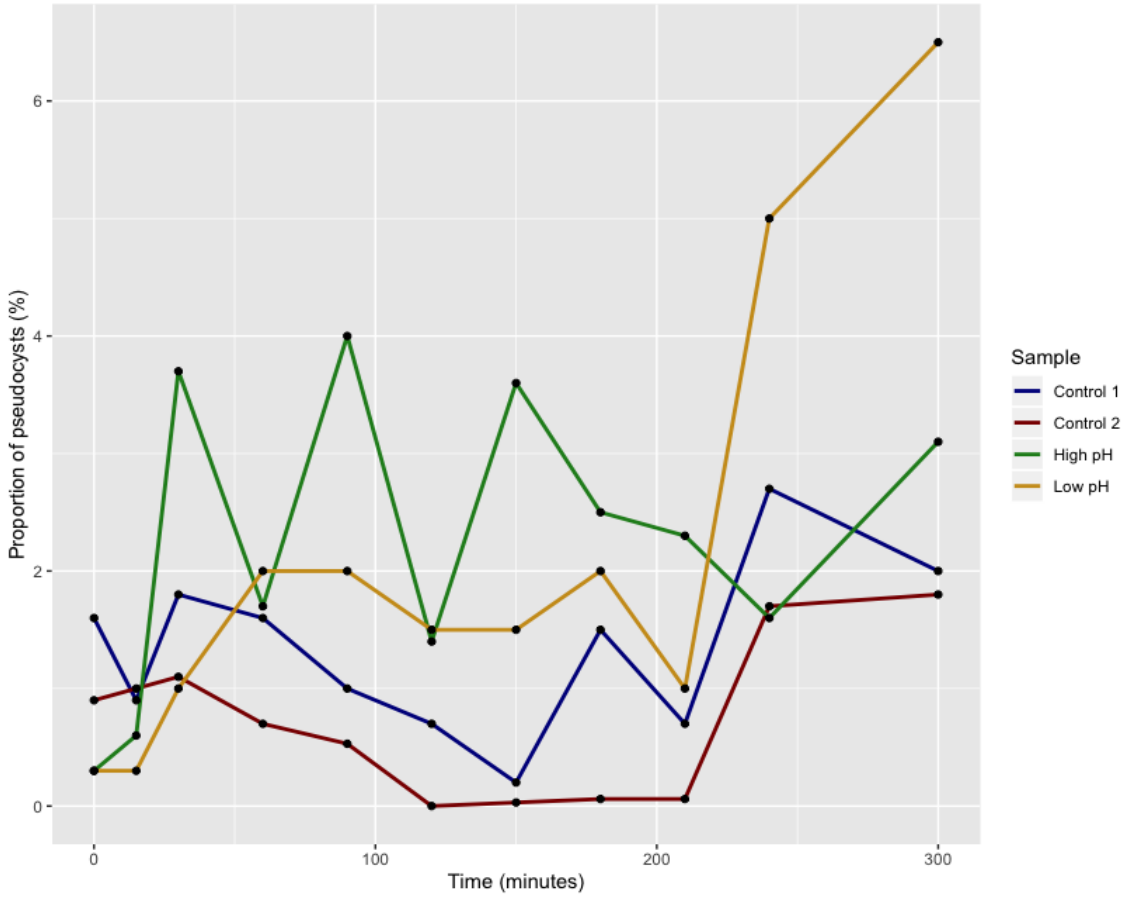

Figure 2.10: Proportion of *T. foetus* cells grown in Diamond media that are in pseudocyst form rather than trophozoite form (controls) when the pH is raised to 8-9 (high pH) or lowered to 5-6 (low pH). Samples were taken initially every 15 minutes for the first 30 minutes (time points 1-3), every 30 minutes from 30 to 210 minutes (time points 3-7) and subsequently every hour from 210 minutes to 300 minutes (time points 7-9). Morphology was identified using light microscopy techniques [150] [16].

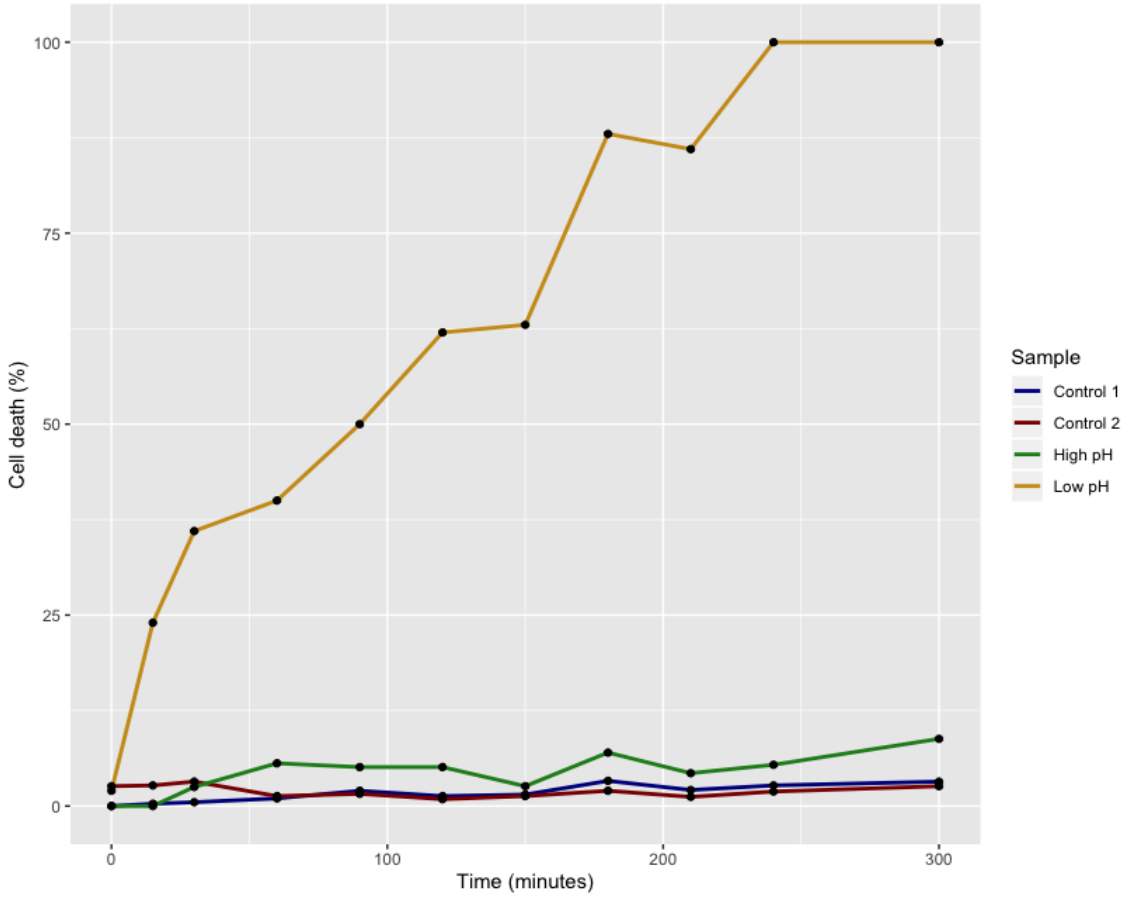

Figure 2.11: Proportion of dead *T. foetus* cells in Diamond media after the pH is raised to 8-9 (high pH) or lowered to 5-6 (low pH) compared to controls in Diamond media with a pH of 7.3 (controls). The pH is changed at time point 0. Samples were taken initially every 15 minutes for the first 30 minutes (time points 1-3), every 30 minutes from 30 to 210 minutes (time points 3-7) and subsequently every hour from 210 minutes to 300 minutes (time points 7-9). Cell death was identified by way of a trypan blue exclusion assay and light microscopy.

When the cells were subjected to high temperature stress the pseudocyst proportions differed again from all other environmental conditions. For the 46°C sample, the proportion of pseudocysts peaks and then falls dramatically and at the same time the cell death increases, suggesting that the cells do not have time to adapt and produce the protective pseudocyst. Whereas, for the 42°C sample, as the temperature increase is less dramatic the cells have time to adapt and form pseudocysts.

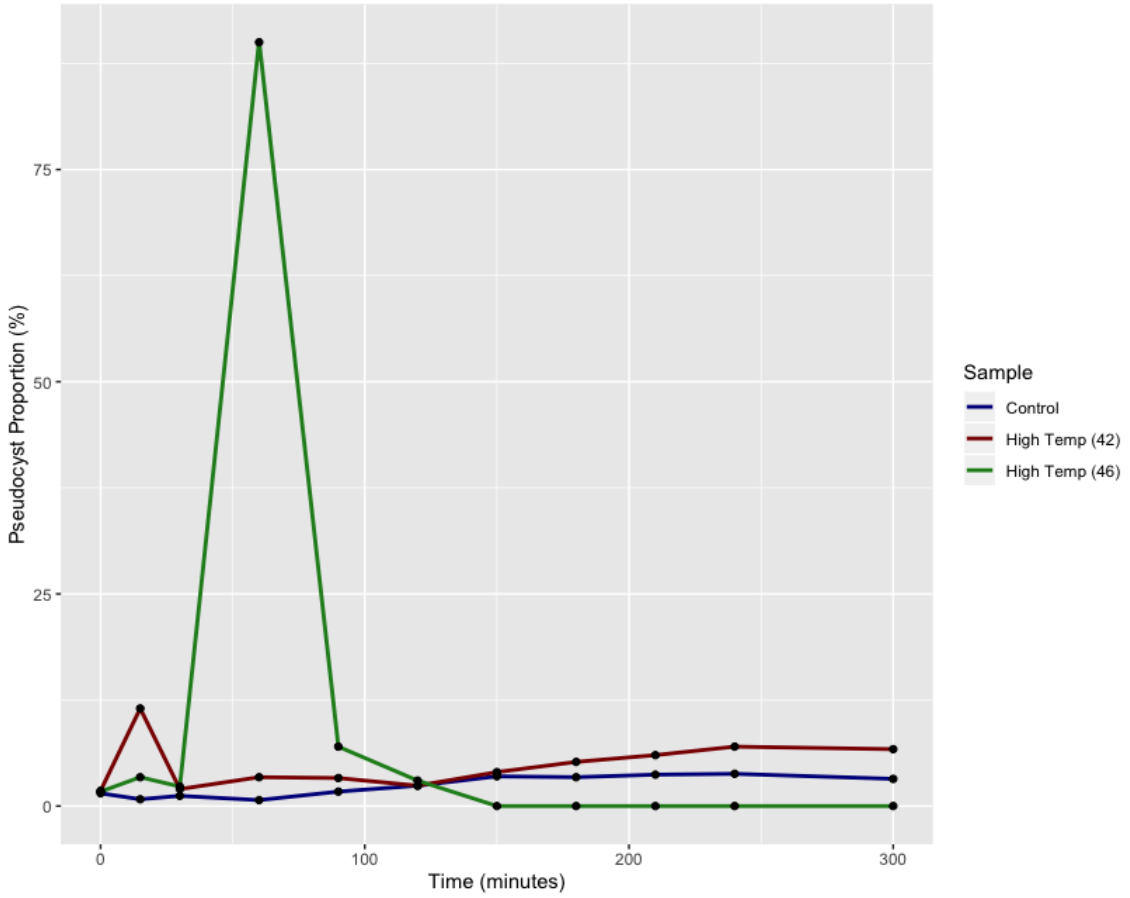

Figure 2.12: Proportion of *T. foetus* cells in pseudocyst form, rather than trophozoite form, in Diamond media after the temperature is raised from 37°C to either 42°C or 46°C. The control cells were continuously grown at 37°C in Diamond media. Morphology was identified using light microscopy techniques [150] [16].

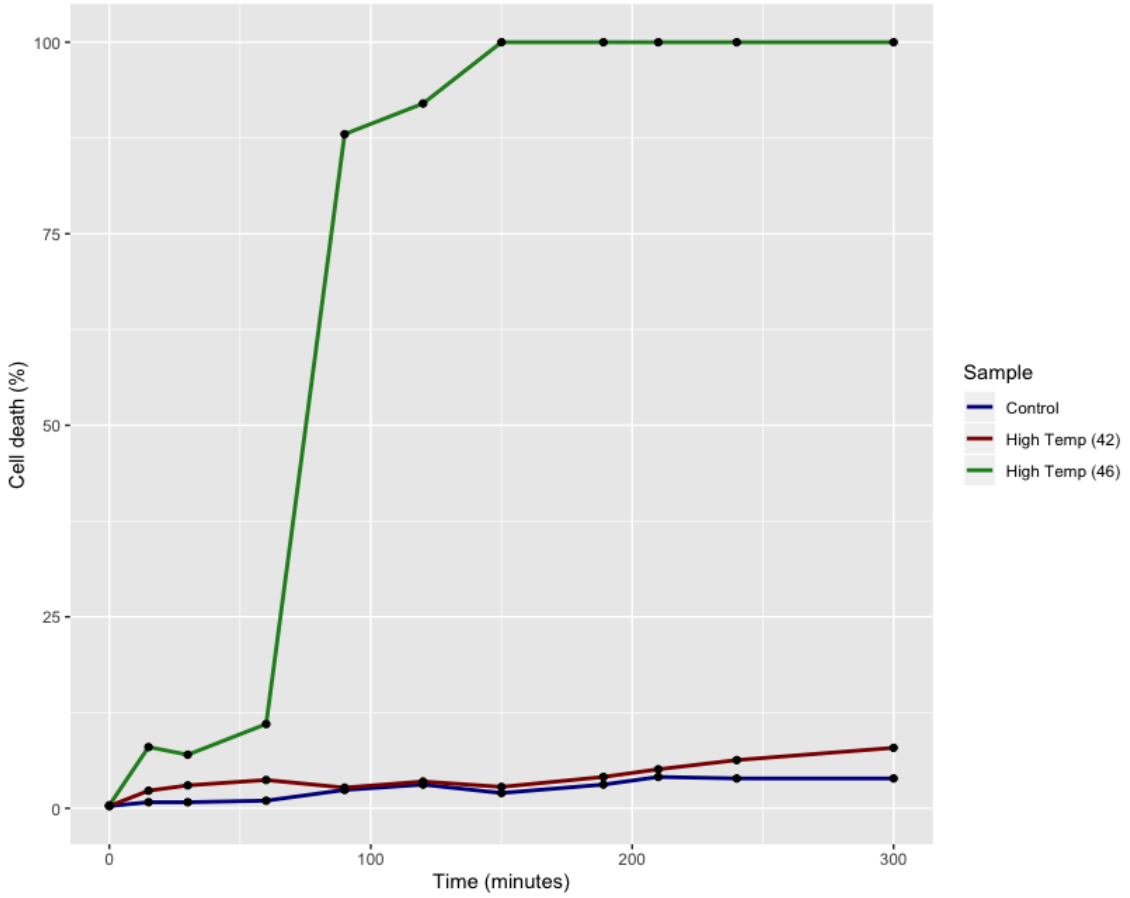

Figure 2.13: Proportion of dead *T. foetus* cells in media after the temperature is raised from 37°C to either 42°C or 46°C. The control cells were continuously grown at 37°C in Diamond media. Cell death was identified by way of a trypan blue exclusion assay and light microscopy.

2.3.6 RNA Seq mapping

All transcriptomes produced were mapped to the genome using Hisat2, there was a wide variation of read number amongst the samples, ranging from 14 million to 67 million reads, but the overall alignment rate was over 90% in all cases (Table A1) apart from Colchicine 2. This was removed from further analysis as it was deemed to have failed in the mapping stage and was an outlier.

2.3.7 Differential Gene Expression

The counts tables produced by Hisat2 and FeatureCounts on the Galaxy Server were used along with the DeSeq2 package in R in order to identify which genes were significantly preferentially expressed in each environmental condition relative to the control trophozoites. For all experiments, control samples are trophozoites grown in Diamond media and incubated at 37°C.

Low temperature

As previously mentioned, low temperature stress is a well-known mechanism for inducing pseudocyst formation in *T. foetus* [262]. The expression of genes within the pseudocysts appears to be more uniform, with a much smaller variance compared to the control trophozoites (Figure 2.14) and all the other transcriptome samples (Figure 2.15). This may be because all cells in the low temperature sample are in their pseudocyst form and have lower metabolic activity, whereas those in the control are a mixed population comprising predominantly trophozoites but also pseudocysts and can be at different stages of the cell cycle.

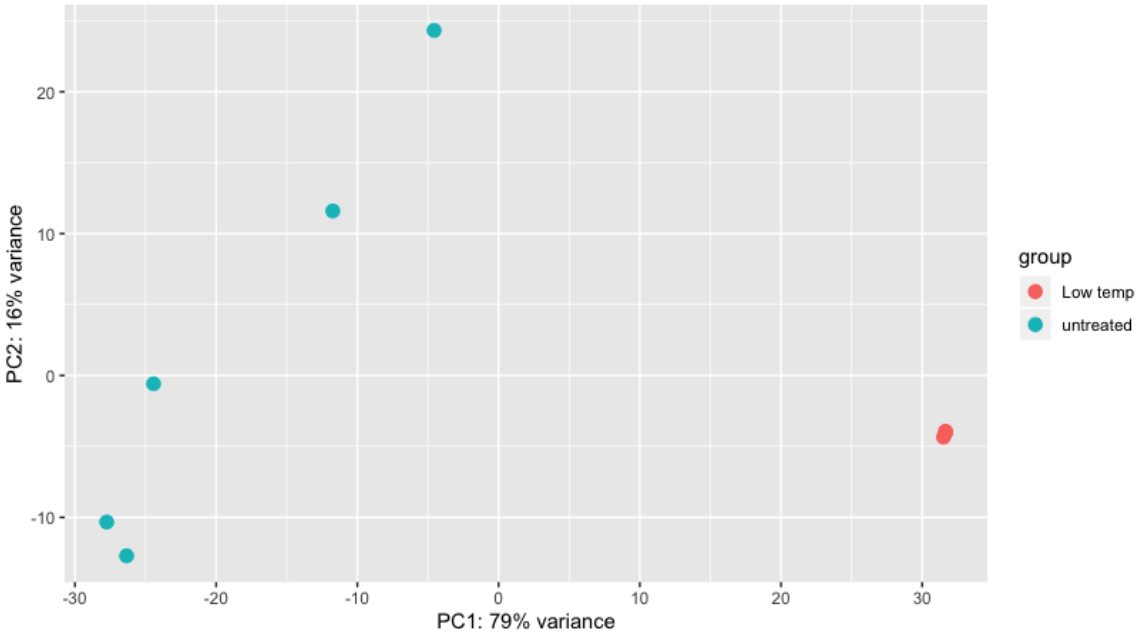

Figure 2.14: PCA plot produced in DESeq2 using principle components 1 and 2. Where untreated (blue) is the control trophozoite sample and treated (red) is the low temperature induced pseudocysts. Pseudocyst morphology was induced by a temperature decrease to 4°C whilst untreated cells were kept at 37°C

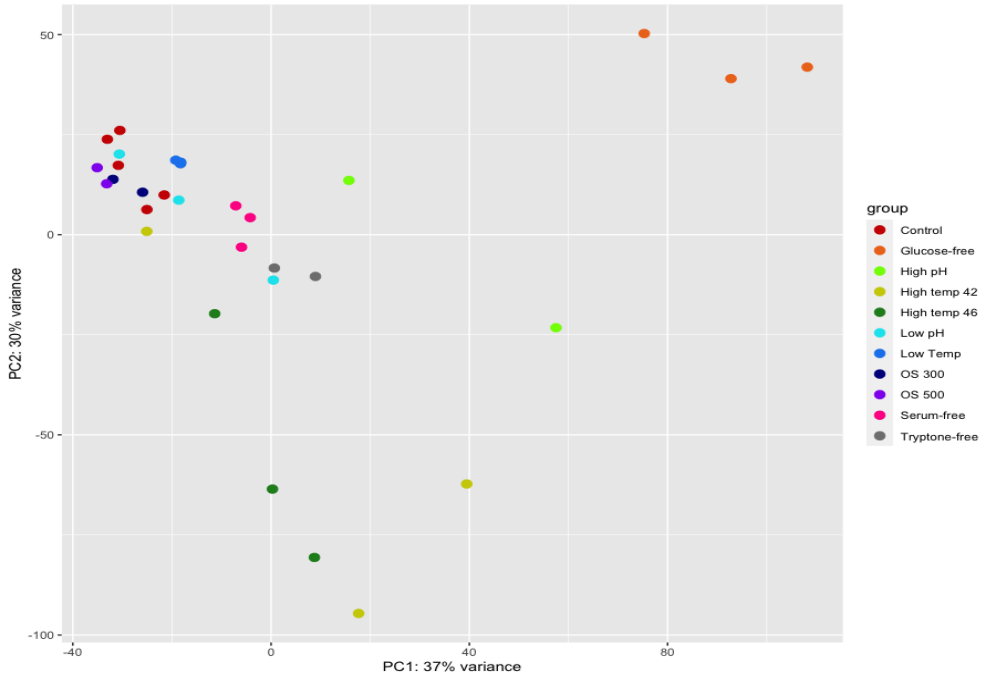

Figure 2.15: PCA plot produced in DESeq2 using principle components 1 and 2 for all transcriptome samples. Where red are the controls, orange are glucose-free samples, light green are high pH, chartreuse are high temperature (42°C), dark green are high temperature (46°C), turquoise are low pH, blue are low temperature, dark blue are oxidative stress (300μM), purple are oxidative stress (500μM), pink are serum-free samples and grey are tryptone-free samples.

There are 1763 genes that are significantly differentially expressed (Figure 2.16) with fold changes ranging from approximately -10 to 15. The most highly differentially expressed genes that are more abundant in the low temperature stress samples relative to the control samples (Table A2) appear to be mainly hypothetical proteins with the exception of two which are coding for N-acetyltransferases.

The downregulated genes in the low temperature samples are more varied, including Myb-like DNA binding domains, which are involved in transcription. This would be expected to be downregulated in the low temperature as the metabolic activity of the cells is reduced, leading to a reduction in gene transcription. Flavodoxin, similarly, is involved in electron transfer so it would be more highly expressed in the control samples with higher activity and lower in the low temperature pseudocysts. The similarities in some of the gene numbers could suggest a repetitive region of genes or exons of a single gene that are currently annotated as separate genes. Upon visualisation of the genome annotation in Artemis it would seem plausible that TTF74316, TTF74317 and TTF74318 are large ORFs that are exons of one very large gene of 4,476 bases. Similarly, TTF74053-4 may be very large exons in a 3,864 base gene.

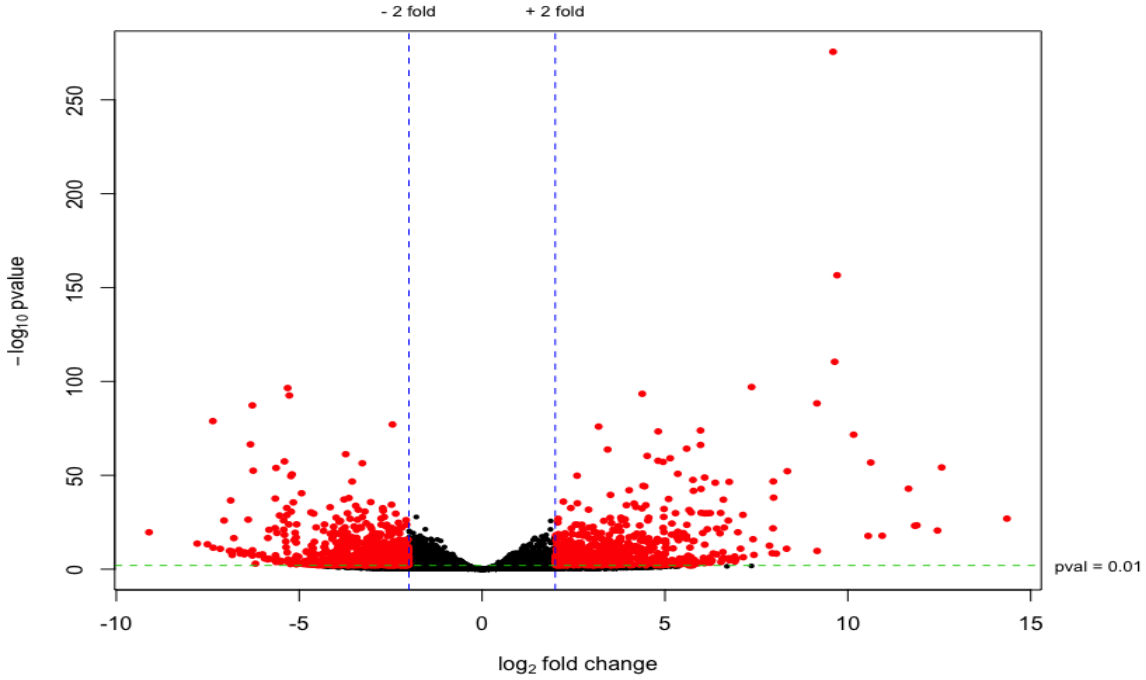

Figure 2.16: Volcano plot produced in R using DeSeq2 showing the differentially expressed genes in the low temperature *T. foetus* pseudocysts compared to the trophozoites. Those significantly differentially expressed (p>0.01 and fold change greater than 2) are shown in red

High Temperature Stress

In the high temperature stress sample, upregulated genes included: ATPases, transferases, many hypotheticals and ABC transporters for both 42°C and 46°C samples (Tables A3 and A4). Whilst downregulated genes included cysteine proteases and calcium kinases. Some of these may be due to the stress response however, very few pseudocysts were produced in the 42°C sample so it is unlikely that the upregulated genes are pseudocyst specific. CAMK family proteins and cysteine proteases are downregulated at high temperatures, it has been postulated that these proteins are involved in the virulence of trichomonads [316] and so increasing the temperature may affect key pathogenicity pathways.

The high temperature stress samples have a higher variance than the controls (Figure 2.15). The 42°C samples have the widest spread, however, both show much wider variance than the other environmental conditions which may be due to the cells breaking down under the stress, particularly as there are several differences between the two temperatures.

Oxidative Stress

The genes upregulated during oxidative stress appear to have similar functions to those upregulated in other environmental conditions, which may suggest that these are the general stress response genes in *T. foetus* (Tables A5 and A6). When 300μM H_2O_2 was used, upregulated genes included; protein phosphatase and CAMK family. This may mean there is increased virulence or a change in biochemical pathways. CAMK itself is a known stress response protein [317]. There are many hypothetical proteins up and down regulated in both 300μM and 500μM oxidative stress responses, these may be *T. foetus* specific stress response proteins.

There is a smaller variance between the oxidative stress samples than most of the other environmental conditions but a larger variance than that seen in the low temperature experiment (Figure 2.15). Relative to the other environmental samples, the oxidative stress appear to have a smaller variance than the controls themselves.

pH Stress

As previously mentioned, *T. foetus* is likely to come under pH stress whilst in the host, particularly if there are changes to the pregnancy status of the cow. For high pH stress, upregulated genes included:phosphatase, transferase, and DNA binding domain GNAt family proteins, which are involved in acetylation (Tables A7 and A8).

Downregulated genes include myb-like domains so there is reduction in DNA binding and a decrease in CAMK family proteins (Tables A7 and A8). It may be that under environmental stress the genes involved in host-parasite interactions have lower activity.

As the cells die more quickly under low pH stress compared to high pH, the reduction in phosphatases may be indicative of the cells reducing their downstream signalling pathways and decreasing metabolic activity, leading to the downregulation of these genes.
The high pH environment samples had a much higher variance than the low pH samples (Figure 2.15). The low pH stress samples had a gene expression profile that was more similar to the control samples than to the high pH stress samples.

Tryptone-free Media

There appears to be a large variance between parasites cultured in tryptone-free media compared to the controls (Figure 2.15). This is likely due to the stress response and utilisation of different

nutrient pathways. However, when compared to the other nutrient depletion samples the variance between the tryptone-free samples and controls appears smaller than the other nutrient-depletion samples, particularly glucose, and the controls and, indeed, each other.

The tryptone-free samples have a much smaller variance relative to the controls than the glucose-free media samples. *T. foetus* may have the ability to utilise different but similar cellular pathways to account for the lack of tryptone in the environment or are using another nutrient in the Diamond media to fulfil a similar purpose. Many genes that are differentially upregulated are skin-secretory xp-like proteins, which are known to have growth factor activity and are secreted (Table A9). Although *T. foetus* does not possess skin itself, it may possess growth factor proteins, the closest homology in terms of primary structure of which is the skin-secretory xp-like protein. The skin-secretory-proteins all have similar gene numbers suggesting that these could be repetitive regions within the genome or are exons of one large gene.

Serum-free Media

The PCA (Figure 2.15) shows that there is a higher variance between the cells grown in serum-free media and the controls. Again this could be due to stress responses. In a similar way to the tryptone-free media exposed cells, genes coding for skin secretory proteins have been upregulated (Table A10). Myb-containing domains were also upregulated suggesting that transcription could be increased under these conditions.

Glucose-free Media

The samples grown in the glucose-free media have a much larger variance than the controls (figure 2.15). When compared to the other media trials there is still a much larger variance with regards to cells grown in glucose-free media compared to the others. This suggests that the lack of glucose in the media promotes a more distinct response within the cells than any of the other conditions. This response is likely to be due to glucose being a major precursor for many metabolic pathways within the cells and so its absence will have far reaching effects. There are specific up and down regulated genes that are not seen in any of the other environmental conditions, such as glycine dehydrogenase and efflux pumps (Table A11), which are likely to be associated with the change of metabolism.

Many enzymes in trichomonads have multiple copies, for example enolases in *T.vaginalis* [316]. During glucose starvation trials for *T. vaginalis* different paralogues of the same genes were up and

down regulated [252] [318]. Some paralogues are favoured under certain conditions or may perform different functions. For example, under glucose reduction stress, six of seven TV hydrogenosomal malic enzyme paralogues were downregulated and some paralogues of phosphoglycerate mutase were upregulated [318]. *T. foetus* paralogues may perform in the same way with some becoming upregulated in environmental conditions whereas others are downregulated. RNA-Seq and differential expression analysis will identify those with the highest fold change so the expression of all paralogues may not identified as significant. Glucose restriction in other organisms, such as malaria, can also lead to upregulation of virulence genes [318].

2.3.8 Number of Transcribed Genes

For the control samples there were, on average 33,700 genes transcribed which left 50,780 genes non-transcribed. Averages were taken across the samples for each condition, using the feature counts tables, to see how many 'extra' genes were transcribed in the condition. That is, how many genes were transcribed on average per condition, that were not transcribed in the original control samples. Across all conditions, the number of genes that remained non-transcribed ranged from 37,800 to 46,400.

Conditions	Extra Genes Transcribed
Low Temperature	1962
High Temperature (42 ° C)	1394
High Temperature (46 ° C)	1994
High pH	1673
Low pH	736
Oxidative stress (300mm)	457
Oxidative stress (500mm)	453
Glucose free	8935
Serum free	9036
Tryptone free	4737

Table 2.3: The number of additional genes found when compared to control trophozoites across a range of environmental conditions.

There was a much higher number of extra genes found in the samples where a nutrient has been depleted. This could be due to the cells having to utilise different metabolic and biochemical pathways in order to survive, therefore turning on a large range of different genes. However, there

is still a very large number of ORFs that are not expressed, even under these conditions. In serum-free media, glucose-free media and tryptone-free media there are 45154, 48070 and 48822 not expressed respectively. It may be that they require the presence of the host cells to express these genes or they need a different environmental condition.

2.3.9 RNA- GhostKOALA Mapping

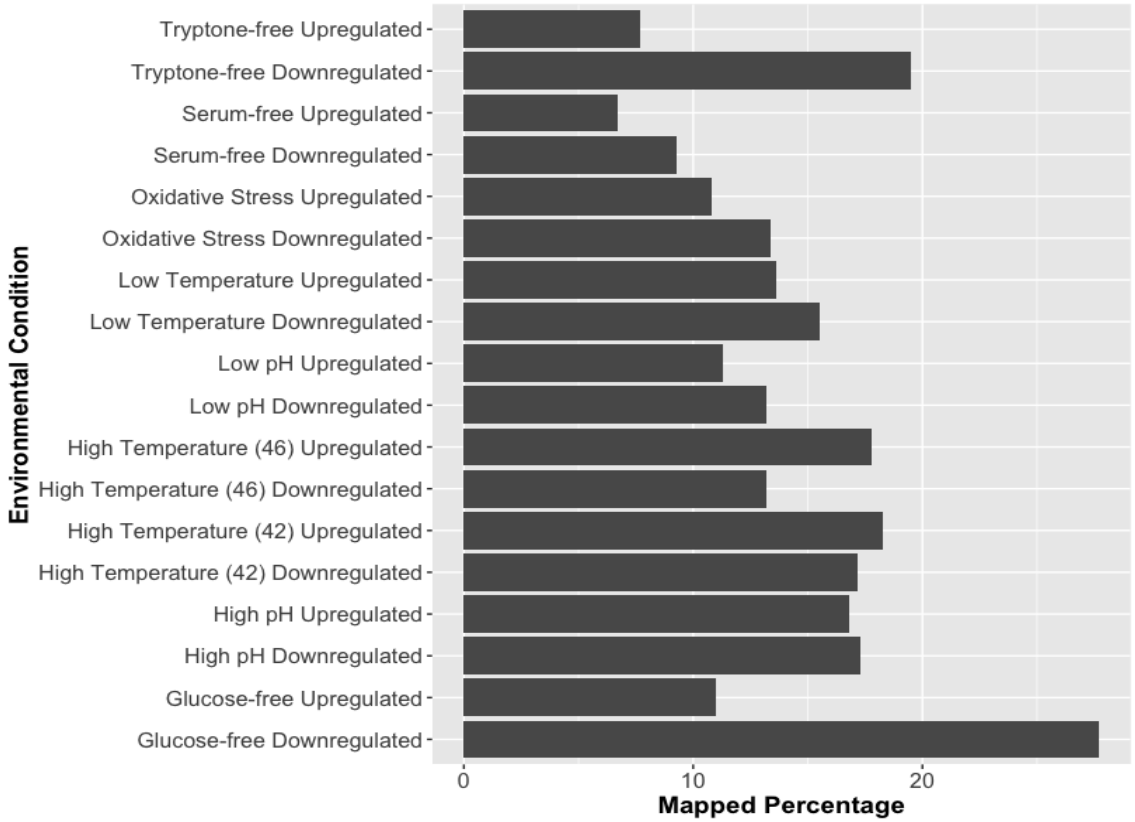

Figure 2.17: Mapping percentages of significantly upregulated and downregulated *T. foetus* RNA transcripts against KEGG pathways. The significantly deferentially expressed genes were identified using DeSEQ2 and mapping against KEGG pathways was performed using GhostKOALA.

When the significantly differentially upregulated and downregulated genes from each environmental condition were run through GhostKoala [288] varying percentages of each group were successfully mapped onto a function. This ranged from over 25% in the case of the down-regulated genes from the glucose-free environment to less than 10% for the serum-free upregulated genes (Figure 2.17). The upregulated sets had lower mapping percentages compared to the downregulated sets in all environmental conditions apart from high temperature stress samples and, overall, the mapping was low, never exceeding 30%.

2.3.10 Comparison of *T. foetus* Genomes

During the course of this project, two more *T. foetus* genomes have been produced [227] [228] [16]. These were all Illumina sequence rather than PacBio. Comparisons were made between my *T. foetus* genome and those produced by Benchimol [227] and Carlton [239] (Table 2.4). The Benchimol genome was assembled using Allpaths whereas the Carlton genome was assembled using Velvet, a *de novo* assembler designed for short reads [240] [241]. When these genomes were compared, my assembly had fewer contigs, a higher N50 and was more in keeping with the size of the *T. vaginalis* genome, Zubacova's [114] predictions and the predictions by SMRT Portal and Jellyfish.

BUSCO [226] [319] was used to measure completeness of the genomes and was also compared with other examples, e.g. *Giardia*. All the *T. foetus* genomes have a similar level of completeness and are comparable with that of *T. vaginalis* (Figure 2.4).

Genome	*T. foetus*			*T. vaginalis*
	Senior	Benchimol	Carlton	Carlton
Isolate/Strain	Belfast	K	KV-1	G3
Data	PacBio RSII	Illumina	Illumina	Sanger
Assembly tool	HGAP2	Allpaths	Velvet	Celera
Size (Mb)	147	65	67	160
Contigs	2776	3730	194,695	74351
Protein coding Genes	84,706	25,030	ND	74,351
N50 (bp)	82,179	34,827	2,054	68,338
% GC (Genome)	30.83	-	-	32.7
% GC (Coding	32.7	-	-	35.5
Spliced genes	17,620	-	-	65
Non-spliced genes	67,086	-	-	59,616
Gene density (/kb)	0.576	0.385	-	0.464
BUSCO score (Complete)	198	203	217	210
BUSCO score (Missing or Fragmented)	105	100	86	93
BUSCO score (Complete %)	65	67	70	69
tRNAs	251	306	ND	468
rRNAs	101	-	ND	668

Table 2.4: Comparison of some of the key features of *T. foetus* genome assemblies, produced by Senior, Benchimol and Carlton, and *T. vaginalis*

The BUSCO scores could be lower than expected because generic eukaryotes were searched rather than trichomonads as there is no trichomonad data set. Also, as *T. foetus* and *T. vaginalis* are parasites, they may be using a reduced number of genes. The differences in BUSCO score could also be due to the differences in assembly methods as there is a large disparity between the numbers of contigs and N50 which are known to have an impact on the overall score [320] and could be due to assembly artefacts.

2.3.11 Comparison of *T. foetus* with *T. vaginalis*

When the assemblies are compared, there are similarities between my *T. foetus* and *T. vaginalis* genomes. They are a similar size, both being over 140 Mb and appear to have a similar number of core expressed genes. Some differences in gene number may be due to the differences in assembly and sequencing method, especially as the *T. vaginalis* genome was produced ten years earlier than the *T. foetus* genome.

The orthofinder [222] results showed a large number, over two thirds of the *T. foetus* genome, of shared genes between the two species (Figure 2.18). Benchimol [227] also found high similarity between species, 72% of ORFs being similar, in terms of sequence, between species. Benchimol also found a higher number of tRNAs in their assembly, although still a lower number than *T. vaginalis* suggesting there is a number of tRNAs not identified in my genome annotation and also that *T. vaginalis* has a higher number in general. This could be attributed to, in part, to the larger genome size. As *T. vaginalis* is not the most closely related trichomonad to *T. foetus* it stands to reason that a large number of these genes are not only conserved between these species but also between most if not all trichomonads.

2.3.12 *In silico* Cell Surface Proteome

In order to reduce the 84,725 genes in the genome to set of plausible vaccine antigens, the first of several filters to exclude unsuitable genes concerned predicted protein structure. Signal P [321] [322] and TMHMM [290] were used to identify signal peptides and transmembrane domains respectively in predicted amino acid sequences. The presence of a signal peptide would show that the protein is probably localised to the plasma membrane and the presence of a TMD would show that it is embedded in a membrane. Initially, Signal P identified 5,205 proteins with a significant probably of possessing an SP. Of these, TMHMM identified 1,607 proteins with one TMD, and 621 with two TMD. When GPIpred was applied, a further 11 proteins were found that contained an SP and GPI

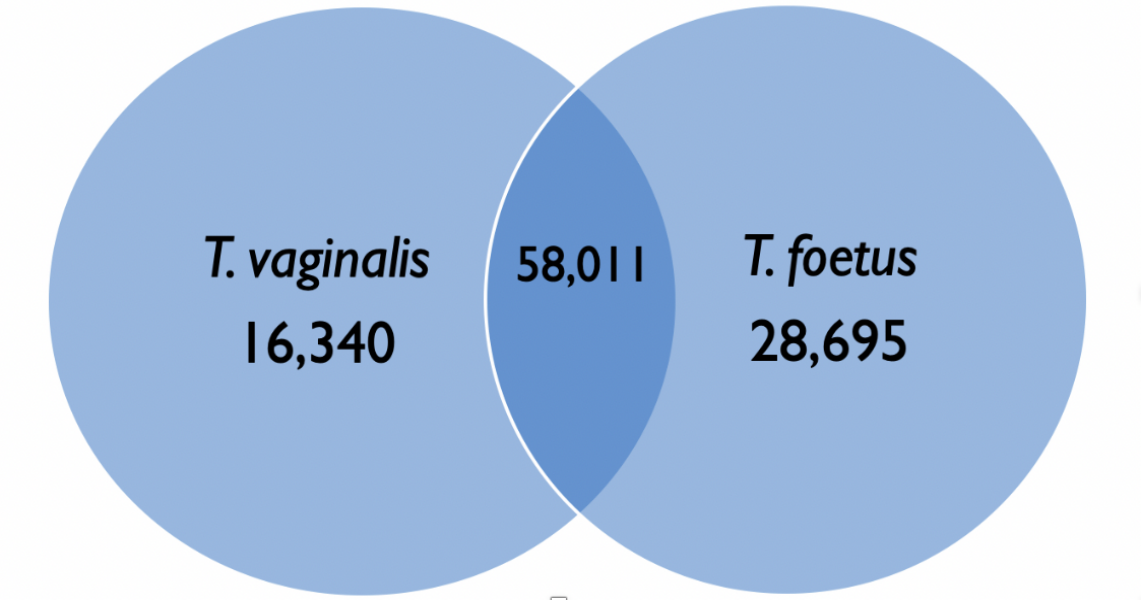

Figure 2.18: Shared genes between *T. vaginalis* and *T. foetus* found by Orthofinder

anchor. Together, this produced a final set of 6,823 genes that are predicted to be expressed on the cell surface, either as transmembrane or tethered proteins, or secreted outside of the cell

2.3.13 Network

All viral proteins and transposable elements were removed from the final protein list, for example, integrases and transposases. Where there were large clusters with only one or two SP positive nodes but many negative nodes, alignments were made and opened in MEGA5 [36] to compare and see whether the positioning of the SP was sensible and likely to be a real result, rather than annotation error. The BLAST scores were also checked to see the length and similarities of the alignments. All families without at least one SP within them were removed, which reduced the sets to 610 families. After the alignments were checked and those with low similarity were removed. This left 246 clusters. Overall there were 105 *T. foetus* specific clusters and 141 clusters with both TV and TF genes within them. The final network contained 11,259 nodes (Figure 2.19).

All *T. foetus* specific genes are coloured in blue (Figure 2.19), all *T.vaginalis* specific genes are coloured in red and all proteins that contain a signal peptide are highlighted in green (Figure 2.20).

Most peptide array genes belong to 'hypothetical' gene clusters however, one belongs to the legume-

like lectin family proteins, one belongs to the Clan-SC family and nine belong to the largest cluster that contains the BspA family, CAMK family and leucine-rich repeats. Very few of the top ten differentially expressed genes from the transcriptome experiments are present within the network, only 21 out of the top 180 differentially expressed genes.

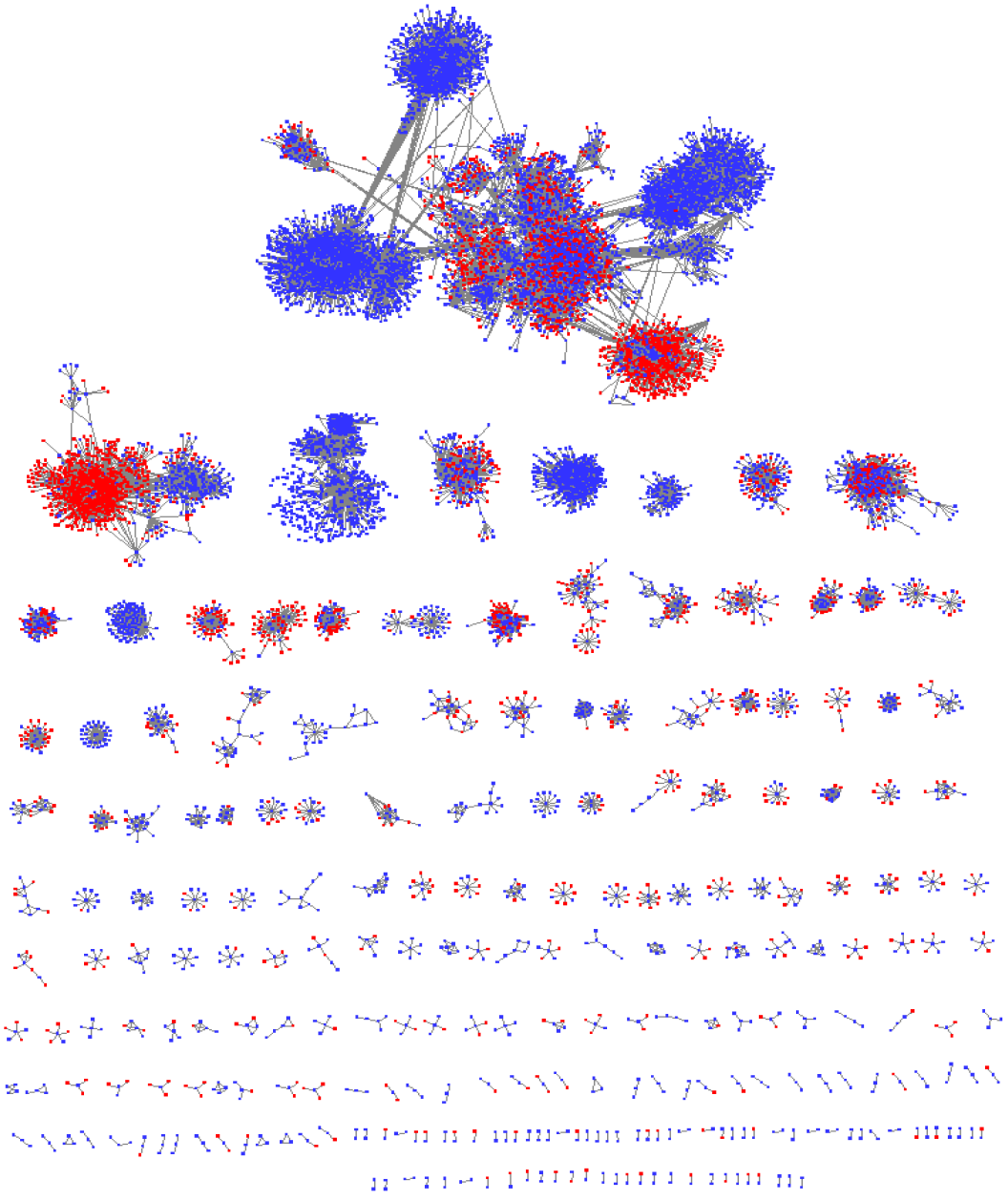

Figure 2.19: Network showing *T. vaginalis* (red) and *T. foetus* (blue) genes produced in Cytoscape. This network shows the *in silico* cell surface predictions of *T. foetus* and their putative gene families based on PSI-BLAST bit scores.

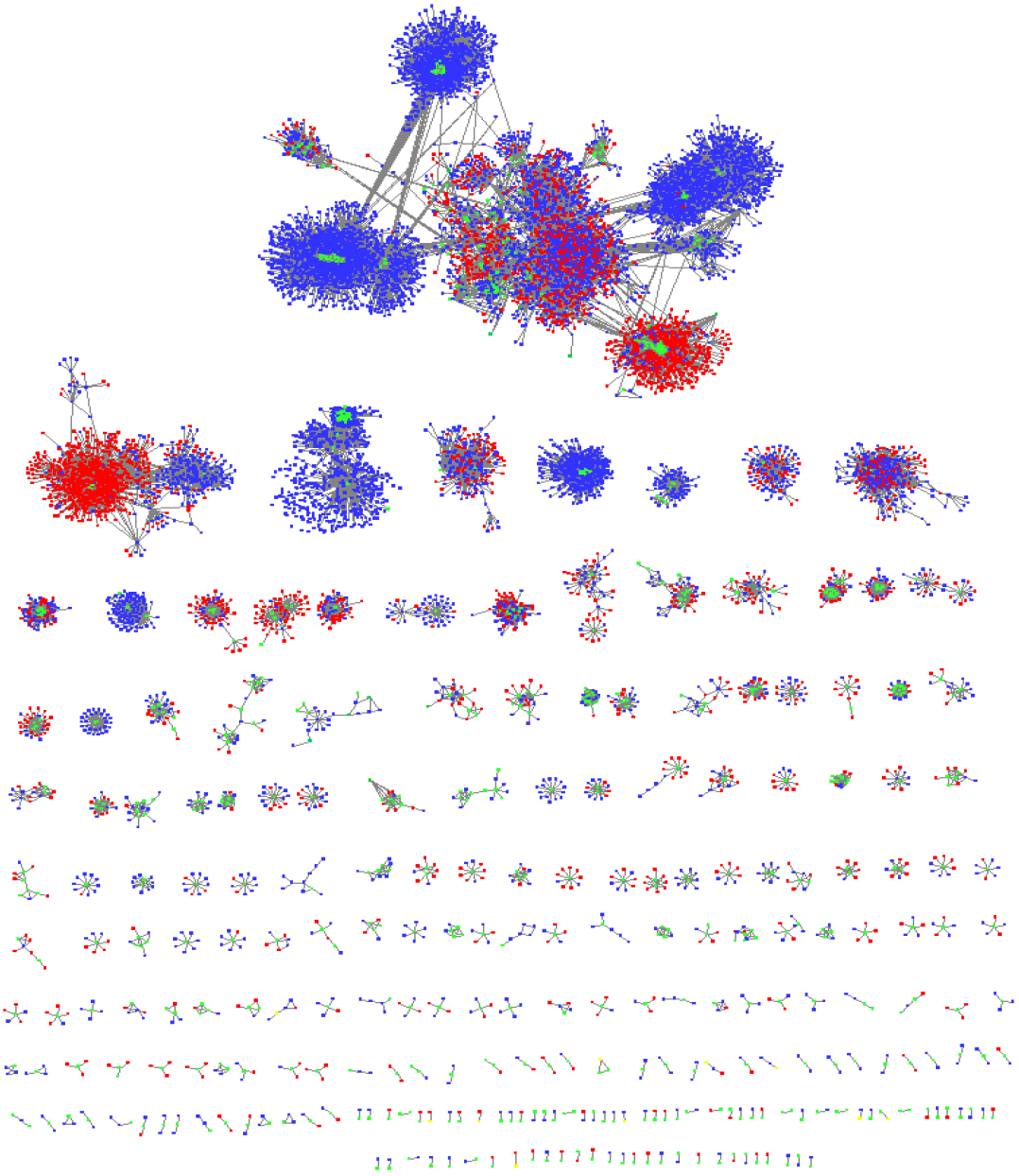

Figure 2.20: Network showing *T. vaginalis* (red) and *T. foetus* (blue) genes, highlighting those that posses signal peptides (green) produced in Cytoscape. This network shows the *in silico* cell surface predictions of *T. foetus* and their putative gene families based on PSI-BLAST bit scores.

Key Gene Clusters

There are over 200 cysteine proteases present in the network, mainly in three clusters. Cysteine proteases are considered to be integral to parasites and their lifestyle [323]. Most parasitic cysteine proteases are contained within the 'Clan CA' family, of which there are 289 in my annotated *T.*

foetus genome. As stated in the introduction, cysteine proteases are involved in cytotoxicity and in *T. vaginalis* the cysteine proteases are known to degrade IgG, IgM and collagen, thereby protecting itself from the immune system and acquiring nutrients from host cells. Within the *D. fragilis* genome, cysteine proteases were the most abundant transcripts found [323] [316]. *T. vaginalis* has 446 proteases, 202 of which are cysteine proteases and 187 of those which belong to the clan CA family [125]. The high numbers of these further suggest the importance of the cysteine proteases to the parasitic lifestyle.

The largest cluster in the network contains: hypotheticals, lecuine rich repeat containing proteins/BspA family proteins and CAMK proteins (Figure 2.21). The BspA family is also thought to be involved in virulence and pathogencity and is highly expanded in *T. vaginalis* [316] [125], it may also be involved in chemotaxis. This family is found amongst many species, both prokaryotic and eukaryotic, the further implications of this are discussed in the chapter discussion.

CAMKs are associated with intracellular calcium and are involved in signalling cascades and regulate gene expression, including likely pathogenic pathways [316]. There are 371 of these CAMKs predicted in the *T. vaginalis* and 71 in the *D. fragilis* genome. Within the present genome, 956 genes have greatest identity to a CAMK family protein. The increase in number compared to the other protists may be due to the multiple hosts and environments that *T. foetus* inhabits. The largest cluster also contains a very large number of kinases (over 1000). Kinases are standard housekeeping genes but, in eukaryotes, they are also linked to pathogenicity [316]. Signal transduction and vesicle trafficking within trichomonads are regulated by these kinases and by Ras GTPases. Kinases may be involved in the necessary cytoskeleton remodelling during host-parasite interactions and during pathogenesis, such as phagocytosis by *T. vaginalis*. The regulation of vesicle trafficking also leads to the regulation of virulence factor secretion, such as cysteine proteases.

The second largest cluster in the network contains ankyrin repeat-containing proteins and hypothetical proteins; some ankyrin genes may have host-pathogen responses, such as that in legionella [324]. This would suggest that the largest gene families in *T. foetus* are involved in host-parasite responses and virulence which is plausible as it is an obligate parasite and its host-interactions are essential for its life-cycle.

There is also a very large 'NA' cluster present, this could be a completely *T. foetus* specific family although the mechanisms of their action and functions are unknown and there is very little structural data on the members. The clusters in the network feature a wide range of proteins including: BspAs, CAMK, Ankyrins, DNA polymerase, Thioredoxin, ATPases and legume like lectins. The

network contains many housekeeping genes along with many hypotheticals which could be *T. foetus* specific gene families. As some do not have any *T. vaginalis* homologs the genes could belong to either *T. foetus* only or more closely related trichomonads that have not yet had their genomes sequenced. Thioredoxins can be found on the cell surface of some protists and have been linked to a protective response against oxidative stresses [325] and may play a role in regulating cysteines.

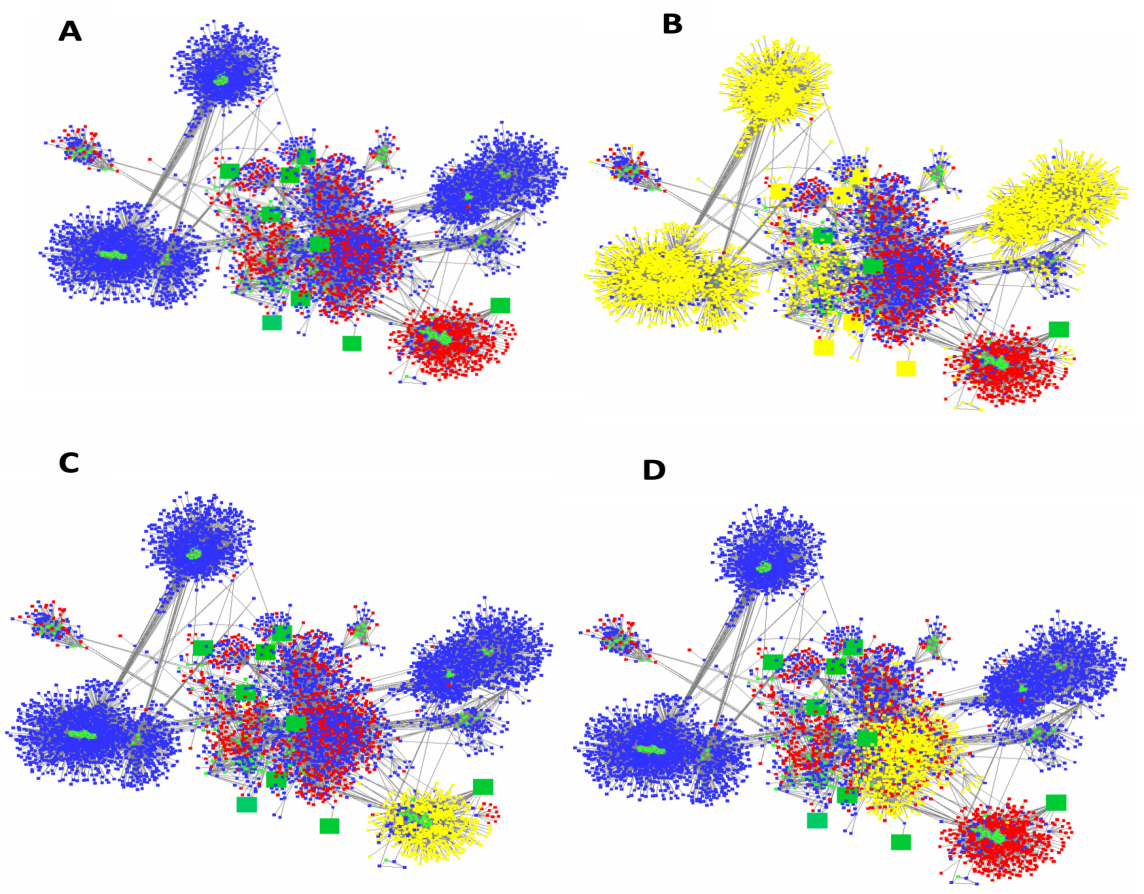

Figure 2.21: Largest Cytoscape cluster (A) highlighted to show: (B) all hypothetical genes, (C) BspA and Lecuine-rich-repeat containing genes, and (D) CAMK family genes. The products for all genes were identified by BLAST2Go using a non-redundant database. To form the clusters PSI-BLAST was used to identify and produce bit scores. Clusters were created in Cytoscape, where *T. foetus* genes are blue, *T. vaginalis* genes are red, signal peptide containing genes are highlighted in green.

2.4 Discussion

In this chapter, a *T. foetus* genome sequence has been assembled and annotated using computational and manual methods. Genome sequencing has been complemented with a transcriptome derived from *T. foetus* cells cultured under diverse environmental conditions, leading to an enhanced genome annotation and to identification of differentially expressed genes under various stresses. Finally, this chapter has produced an *in silico* cell surface proteome of 2,807 genes, containing both conserved and species-specific genes and gene families, which represents the first step in identifying suitable antigens for vaccine candidates.

2.4.1 Evaluation of Genome Assembly

The genome was assembled by a variety of methods, including HGAP2 (SMRT portal) and Canu [271], with Canu predicting a size of genome double that of HGAP2. The size was also predicted using Jellyfish [326][237] and this was later used in conjunction with HGAP2 to produce the final genome size. Canu produced a much more fragmented genome with 11,000 more contigs and this could be due to overexpansion of repetitive regions. Also, previous comparisons that have been performed between assemblers are in bacteria rather the trichomonds or other eukaryotes [326]. The size of the SMRT Portal assembly seems more inkeeping with the size predicted by Jellyfish and Zubakova [114] and is a similar size to *T. vaginalis*. The *T. foetus* genome is known to be highly expanded and to contain a large number of duplicated genes [316].

2.4.2 Comparison of *T. foetus* Genomes

The present PacBio-based genome assembly is twice as large, and contains more genes, than the other *T. foetus* genome assemblies that have been created previously. Furthermore, the present assembly is more consistent with predicted sizes of the genomes from a variety of sources. There are also differences in N50 and contig numbers between the genomes though it is hard to get a direct comparison due to differences in assembly and sequencing methods. When the *T. foetus* [227] and *T. vaginalis* genomes were initially compared by Westrop [242], it was shown that 72% of ORFs were similar between species and 28% were found in *T. foetus* only. This is similar to the orthofinder analysis performed in this chapter, which shows that 66% of gene clusters were shared between species and 34% were found in *T. foetus* only. The differences in percentages could be attributed to the fact that the Benchimol genome was much smaller (65Mb opposed to 147Mb) and

had significantly fewer ORFS (25,000 opposed to 84,700), therefore fewer *T. foetus* specific genes may be found.

2.4.3 Comparison of the *T. foetus* and *T. vaginalis* Genomes

The *T. vaginalis* genome is predicted to be larger than the *T. foetus* genome [20] and the present assembly conforms to this, being 20Mb smaller than that of *T. vaginalis*.
T. vaginalis secretes large amounts of cysteine proteases and exosomes, the secretome of *T. foetus* would also be interesting to investigate to see if the same repertoire of proteins are secreted between the two species. Looking at the network of putative cell surface proteins, there is a large cysteine protease family of genes, each possessing a signal peptide and no transmembrane domain. This shows that secreted cysteine proteases are also produced by *T. foetus* and perhaps fulfil the same functions.

Screening the *T. vaginalis* genome and proteomic analyses identified many BspA and Pmp family proteins[20]. BspAs are leucine-rich repeat containing proteins that are found in both prokaryotes and eukaryotes. Pmps are known to promote binding, making them a potential virulence factor [167] [327]. As BspAs are found in prokarotes, it is thought that they were introduced to eukaryotic species via horizontal gene transfer. *T. vaginalis* is thought to have 911 BspA proteins and 48 Pmp proteins encoded in its genome [165]. The present *T. foetus* annotation includes 186 proteins that are annotated as BspA or leucine rich repeat proteins and 12 Pmp labelled proteins. Although these numbers are significantly lower than that of the *T. vaginalis* values, they are consistent with previous findings. Handrich *et al.* [327] compared the gene expression of BspAs and Pmps in trichomonads to general housekeeping genes and found a large expansion of these families in *T. vaginalis*. The relative abundance of BspAs and Pmps compared to housekeeping genes was 506 and 80 respectively. Whereas in *T. gallinae* it was 47 and 6, in *T. tenax* the abundance was 81 for BspA and 11 for pmps and in *Tritrichomonas batrachorum* it was 295 and 2 for relative abundance. *T. foetus* was not included in the study.

There is also a large family of rab-GTPases which also appear in the *T. foetus* genome. These are involved in controlling intracellular membrane traffic and ensure membrane bound proteins are located to the correct site [328]. Additionally, these have also been postulated to be involved in pathogenic processes such as the secretion of virulence factors and cellular motility [316] [329]. There are over 65 Ras and Rab GTPases within the *T. vaginalis* genome and a similar number within *T. foetus*- 49 predicted Ras GTPases and 18 predicted Rab GTPases. The large number of

CAMK proteins in both genomes also suggests a key role in the parasites lifestyle.

The BspA family of proteins appears to be key in both *T. foetus* and *T. vaginalis* [167] and is likely involved in host-parasite attachment. It is known to be involved in surface adhesion in many bacteriodes [167]. It is likely that for many parasite-host and virulence proteins in *T. vaginalis* there is a *T. foetus* homologue or protein that performs a similar function. These genes may provide key vaccine targets for *T. foetus*, as knocking them out could greatly reduce the virulence and pathogenicity of the organism. Only 65 genes in the *T. vaginalis* genome have introns compared to 18,000 in the *T. foetus* genome [125]. This may be due to *T. foetus* colonising many different environments, such as the urogentital tract and gastrointestinal tracts [30] [330], thereby needing a higher number of potential proteins. This could also be due to misannotation of either the *T. foetus* or *T. vaginalis* genomes. There is currently no other trichomonad genome that has been sequenced using PacBio long-read technology and very few trichmonad genomes that have been sequenced by any technologies. The genes that are found in both *T. foetus* and *T. vaginalis* could be conserved between all trichomonad species, for example: CAMK and rab GTPases.

Additionally, there is a large difference between the genomes in the number of intron containing genes, with *T. vaginalis* containing 65 genes and *T. foetus* containing 17,602. although some protists are known to have relatively few, many, such as *Plasmodium falciparum* have thousands, suggesting that the number found in *T. foetus* are still inkeeping with the overall intron densities for protists.

2.4.4 Evaluation of Gene Annotations

In the present assembly, 12,106 genes were defined as 'NA's and a further 44,953 were annotated as 'Hypotheticals'. These proteins are not necessarily species-specific, but they are uncharacterized, perhaps because they have limited taxonomic distribution among parabasalids. Few trichomonad genomes have been thoroughly sequenced beyond the model *T. vaginalis* and, therefore, both sequence and functional databases will not represent much of the core trichomonad gene set, Thus, while 56,000 genes are found in both *T. foetus* and *T. vaginalis* genomes, these have no affinity beyond trichomonads and no ascribable function.

2.4.5 GhostKOALA Mapping

When the genes were mapped to KEGG pathways using GhostKOALA, many pathways were not mapped to. Furthermore, of the pathways that were mapped to, many did not have complete modules,with some only containing one member. Only 18 modules in total were fully mapped

to and only 11.6% of genes in the genome were successfully mapped to a pathway. This may be due to a number of reasons: misannotation in the original genome; the parasite could be using a reduced set a pathways due to its parasitic lifestyle and is gaining some of its required nutrients and compounds from the host; or, could be due to the fact that the trichomonad genes are too divergent and so GhostKOALA does not recognise them.

2.4.6 Errors in Genome Annotation

Eukaryotic genes typically consist of multiple exons and introns, which can make correct gene modelling more challenging [331]. Therefore, automatic gene finders are not always accurate. For example, when the *Mycoplasma genitalium* was sequenced [332] the error rate was found to be at least 8%. Genes with incorrect sequence annotation can be entered into public databases which leads to further annotation errors down the line with other genomes. In order to try to combat this inaccuracy, I used two automatic gene finders along with manual curation to minimise obvious errors. One of these gene finders, BRAKER, also implements RNA-Seq data to identify ORFs and allows for small genes to be found [331]. Gene finders in general, however, follow a set of assumptions which, in some cases may not always be correct or valid [230]. These include:

1) Start and stop codons and splice sites are the same throughout the genome. This is particularly a problem when annotating a non-standard organism such as *T. foetus* when the true splice patterns and codon usage are unknown.

2) No genes in the genome overlap.

3) No genes are nested within other genes.

4) The lengths of the introns and exons are geometrically distributed along the genome. AUGUSTUS is known to be an exception for this

5) There are no sequencing errors in the input sequence. PacBio and other long-read technology are known to have a relatively high error rate.

6) There is no alternative splicing. Many eukaryotic genes use alternative splicing to increase the number of proteins that can be coded for. If alternative splicing is not taken into account many genes could be missed. Furthermore, it may not be the highest scoring isoform that is always predicted by the programs, leading to errors in the gene sequence or functions.

7) The start and stop introns contain no genes.

To attempt to overcome these problems, as previously stated, a range of gene finders were used along side manual curation with extensive transcriptome analysis to minimise the error rates.

2.4.7 Network

The network produced contains many expected protist genes including the TLK family and thioredoxins. The largest cluster in the network features the BspA family and lycine-rich repeat containing proteins, both of which are thought to be key virulence factors in *T. vaginalis* or are involved in host-parasite interactions. As stated in the results, this cluster also contains a large number of kinases which are crucial for a wide range of cellular pathways. There were many hypothetical proteins that did not show homology to any *T. vaginalis* genes, including several of the peptide array genes. In the largest gene cluster it would be reasonable to infer that the hypothetical *T. foetus* genes display similar functions to the other family proteins, such as kinase activity and it is likely that the relationships between the different nodes correlate with functional relationships between genes [264]. Within the *T. vaginalis* hydrogensome there is the thioredoxin-linked periodoxin system that removes reactive oxygen species and can convert H_20_2 to H_20 [325]. The trichomonad thioredoxin reductase differs significantly from that found in humans so may be a good drug target for *T. vaginalis*. If it is found that the *T. foetus* thioredoxin is different from the bovine and feline thioredoxin, which is likely, this too could be a good potential drug target.

2.5 Conclusion

The genome of the Belfast strain of *T. foetus* has now been sequenced and assembled, producing 84,725 protein coding gene models. From among these, a predicted cell surface proteome has been identified containing 2,807 genes, which are likely to be expressed on or beyond the plasma membrane. This gene set can now be validated and tested by immunoproteomic methods to confirm that they fulfil the original criteria for a good vaccine candidate (Figure 2.22).

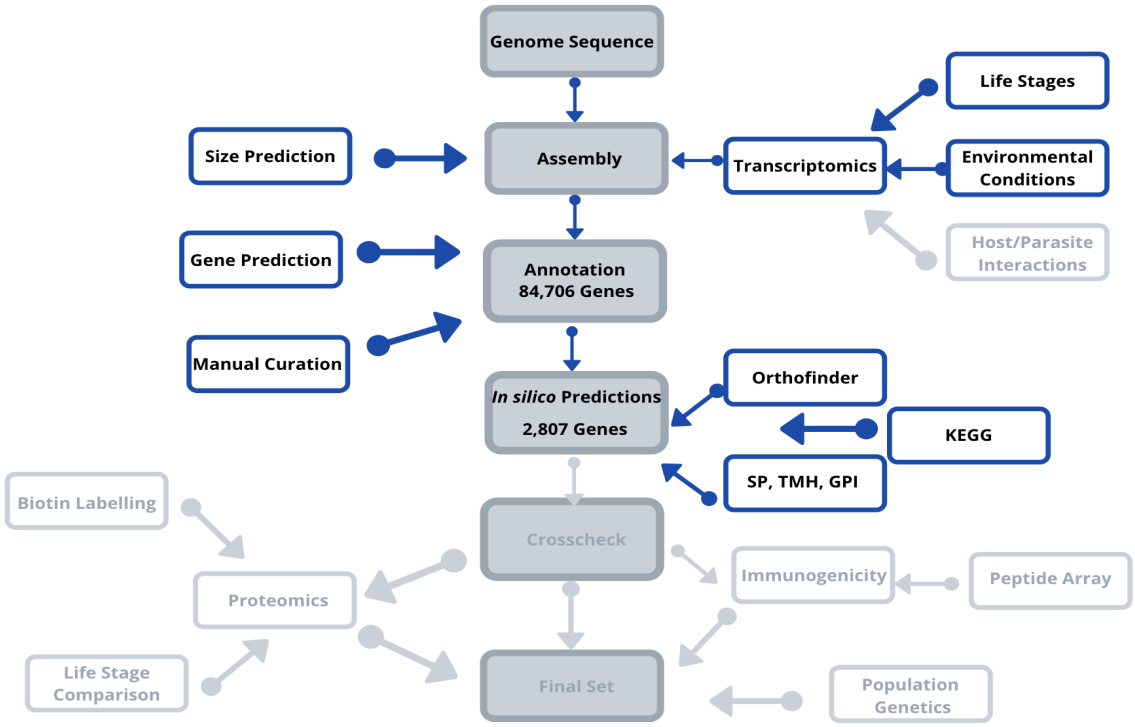

Figure 2.22: Reverse Vaccinology Project Flow Diagram. The sections highlighted are those that have been completed by Chapter 1, other sections are the subject of following chapters.

Chapter 3

Proteomic Analysis of *T. foetus*

3.1 Introduction

This chapter describes proteomic analysis of the *T. foetus* cell surface in two parts. *T. foetus* cells are labelled with biotin and the cell surface proteins are then extracted using a biotin-streptavidin pulldown. These proteins are analysed using Tandem Mass Tag (TMT) labelled mass spectrometry and compared to the predicted cell surface proteome described in Chapter 2 to validate the *in silico* approach. Both trophozoites and pseudocysts are analysed using mass-spectrometry to identify differentially expressed proteins between the two developmental stages. In the second part, a peptide microarray is designed from 50 cell surface proteins identified in Chapter 2, which is then probed with serum from natural and experimental *T. foetus* infections, to examine their immunogenicity. Natural immunogenicity is indicative of good vaccine candidates and will contribute to a short list of antigens that can go on to further stages of vaccine development [333] [334].

3.1.1 Proteomics

Proteomics is the study of all proteins produced by a particular organism, their diversity and relative abundance, and is used to complement other 'omics' methods, such as genomics and transcriptomics [335]. Proteomics approaches can be used to identify biomarkers and vaccine candidates, in addition to providing more information of disease progression [336], aiding drug discovery [335] and identifying key processes in biochemical pathways [337]. The proteome is much more complex than the genome and changes in gene expression can be identified by both transcriptomics or proteomics. The cell proteome can reflect the transcriptome, however, since the amount of protein produced is not solely controlled by the expression level of a particular gene [242], and since many proteins

can be modified, for example by post-translational modifications (PTMs) such as glycosylation, transcriptomes and proteomes can give very different accounts of gene expression. Conventional techniques for protein purification and analysis are ELISAs, Western blots and chromatography [335], however, these do not typically have high-throughput capabilities. Mass spectrometry (M/S) has the ability to handle many samples at the same time and has benefits for a wide range of organisms, including parasites [338]. In the case of *Leishmania*, half of predicted genes are hypothetical (similar to the *T. foetus* genome) [339] which makes purely genomic based approaches for identifying their function difficult. Proteomics has been used to look at many stages, including the salivary components of the *Leishmania* vector, the sand-fly [340], and *Leishmania* secreted proteins [338]. Gene expression is regulated at transcription, translation and by PTMs, therefore a variety of approaches are needed to correctly deduce their roles and functions, this can be particularly important in host-parasite interactions. Expression proteomics can identify whether protein levels are up or down-regulated in different life stages or states, such as trophozoites and pseudocysts in *T. foetus* and comparisons can be made [338].

3.1.2 Cell Surface Proteomics

The cell surface plays key roles in many different aspects of cellular interactions including in infection and disease [341] [342] and can be ideal targets for drug therapies or vaccines [341]. Cell-surface proteomics has been used to investigate many diseases, such as cancers [343]. Labelling the cell surface and performing a proteomic analysis has been used to identify host-parasite interactions, drug resistance and possible environmental adaptations [342].

In *Trypanosoma brucei*, cell-surface proteomics was used to identify stage specific paralogues that may be involved in cellular remodelling and adaptation to the environment [344]. To identify cell surface proteins of *Giardia* [345], cells were labelled with biotin (Section 3.1.8), incubated with streptavidin and were immunoprecipitated by streptavidin affinity chromatography. 86 peptides were identified by liquid chromatography and tandem mass tag (TMT) spectrometry, 15 of which were known *Giardia* cell surface proteins. Comparisons between the biotinylated proteins of two strains of *Giardia* showed 23 highly conserved cell-surface proteins [345]. These conserved proteins could be good vaccine candidates if they were found to validate all other necessary criteria as they could elicit an immune response to several if not all, strains.

3.1.3 Mass Spectrometry

Mass spectrometry measures the mass to charge (m/z) ratio of ions in a sample. Proteins are digested into peptides which are then fractionated. There are several ways to fractionate these peptides, such as: protein digestion on 1D SDS page gel, liquid chromatography or buffer exchange [336]. The mass spectrometer generates multiple ions from the peptide sample under investigation and separates them according to the specific m/z ratio and records the relative abundance of each ion type. The relative abundances are compared to a database of known peptides and the identity of the proteins present in the sample may be inferred computationally. Liquid chromatography mass spectrometry (LC-MS) and matrix assisted laser desorption ionization-time of flight mass spectrometry (MALDITOF-MS) are the two most widely used methods of M/S [335]. The development of high throughput M/S has produced far reaching benefits for the areas of immunology and vaccine design [336]. LC-MS has been used to identify differences in metabolism between species, such as *T. foetus* and *T. vaginalis* [242]. It was also used to compare the proteomes of bovine and feline *T. foetus* genotypes [346], which showed highly conserved proteomes between the genotypes, but with significant differences in the abundance of 25 peptides, particularly cysteine proteases which are thought to be involved in host-parasite interactions and virulence.

3.1.4 Labelled and Unlabelled Mass Spectrometry

Unlabelled or label-free mass spectrometry involves either counting the peak areas for all of the peptides found or counting the spectra produced from a sample [347] [348]. The more abundant the protein, the more peptides that will end up being fragmented. Label-free M/S is less accurate overall but has greater coverage of the proteome and can identify even very low abundance peptides. It is also cheaper and does not rely on good labelling efficiencies [348]. Labelled M/S produces more sequence ions, is usually more accurate and may increase confidence of identifications using the final database [349] [350]. Examples of labelled M/S included labelling with stable isotope, such as SILAC and using isobaric tags, for example in TMT labelling.

Labelled Mass Spectrometry-SILAC

Stable Isotopic Labeling with Amino Acids in Cell Culture (SILAC)- cells are grown in culture with the addition of stable-isotope 'heavy' forms of amino acids. The peptides produced will be labelled with heavy or light forms of these amino acids and are then differentiated through M/S [351]. It is

used predominately to look for changes in cellular pathways and the regulation of PTMs and gene expression [335] [351].

Labelled Mass Spectrometry-TMT

Tandem Mass Tag Labelling (TMT) involves covalently attaching an isobaric tag, which is a stable isotopic label to the different peptide samples [352] [353]. This allows the samples to be differentiated during the M/S detection stage. The tags separate under fragmentation and release the reporter ions. The reporter ions produced are characteristic of each tag and are detected at a distinct m/z ratio, the intensity of these ions is used as a quantitative measurement [352]. An advantage of TMT labelling over using stable isotopes is that there is not the limitation of 2-plex or 3-plex sets of reagents and so many samples can be analysed in one run, thereby reducing time and cost [352]. It also helps to remove sample preparation variability. However, only one peak is produced per peptide, which can lead to lower sensitivity if there is a small number of highly expressed peptides.

3.1.5 Applications of Biotin in Protein Characterisation and Purification

Biotin, also known as vitamin H or vitamin B7, is present in small amounts in all living cells, where it is a key player in gluconeogenesis and fatty acid metabolism. It is found in low amounts in many foods, such as oats, molasses and soybeans [354]. In animals there are several biotin-dependent carboxylases, such as pyruvate-carboxylase and acetyl-co-A-carboxylase-1 [355]. Many biotin molecules can bind to a protein molecule. It is a small molecule (244.3 Da) and can be attached to the amine groups of diverse proteins and molecules without altering the biochemistry.

Biotin has the ability to bind to avidin or streptavidin to form a complex (Figure 3.1) and is widely used in protein purification techniques and identifying protein-protein interactions and PTMs [356]. Biotin can also be modified as needed, such as by the addition of disulphide bridges, extra-long linkers and many different reactive groups, for example, PEG-4. These can make the biotin more soluble and easier to cleave from the reactive groups, thereby reducing the need for harsh elution techniques. There are several reasons why biotin labelling is used so frequently in protein purification, including: the ability to purify under extreme conditions due to the strong covalent bond between biotin and avidin and the high sensitivity and low background noise. Different types of avidin can be used in experiments depending on the specific needs [356]. In this chapter, I attempted

to use these properties of biotin to label *T. foetus* cell surface proteins prior to identification with label-free mass spectrometry.

3.1.6 Avidin and Streptavidin

Avidin is thought to be a natural antibiotic found in the egg whites of reptiles and birds and produces biotin deficiencies in experimental models [355]. It is a tetrameric glycoprotein (67kDa) with four identical subunits and each subunit can bind to a biotin molecule (Figure 3.2). The binding occurs by the linking of avidin to tryptophan residues and lysine on the subunit binding site.
There are, however, some disadvantages of using avidin [357], predominantly its non-specific binding. Avidin is positively charged at neutral pH so may bind to charged cellular proteins and it is a glycoprotein so may react with lectins. The type of avidin used in an experiment can be modified depending on the needs, such as lower binding affinities to assist with elution [356].

Streptavidin is an 52.8kDa tetramer and overcomes some of the limitations of the biotin-avidin method. It is isolated from *Streptomyces avidinii*. It has less positively charged residues, due to having a more neutral isoelectric point (avidin's isoelectric point is closer to 10) therefore will be less likely to non-specifically bind to positively charged cellular structures. It is also not a glycoprotein so does not bind to lectins.

3.1.7 Biotin-Streptavidin Interactions for Protein Purification

Biotin-streptavidin binding is one of the strongest known non-covalent molecular interaction between a protein and a ligand (Figures 3.1 and 3.2) with a Kd value of $1x10^{-14}$ (insert Duan 2012 reference). The bond is formed rapidly and is unaffected by environmental extremes, such as changes in pH and temperature.
However, there are issues with non-specific binding and also the conditions needed to elute the proteins are often harsh which can affect the structure of the proteins that are trying to be captured.

3.1.8 Biotin Labelling and Pulldowns

Biotin labelling of the cell surface or proteins that are expected to localise to the surface has been widely used. It can be used to enrich samples before mass spectrometry, allowing cell surface antigens to be identified [358]. It can be used to detect proteins that are trafficked to the cell membrane [359]. Identifying cell surface proteins by mass spectrometry has always proved challenging [342] [360], particularly plasma membrane proteins. This due to the complexity of cellular membranes

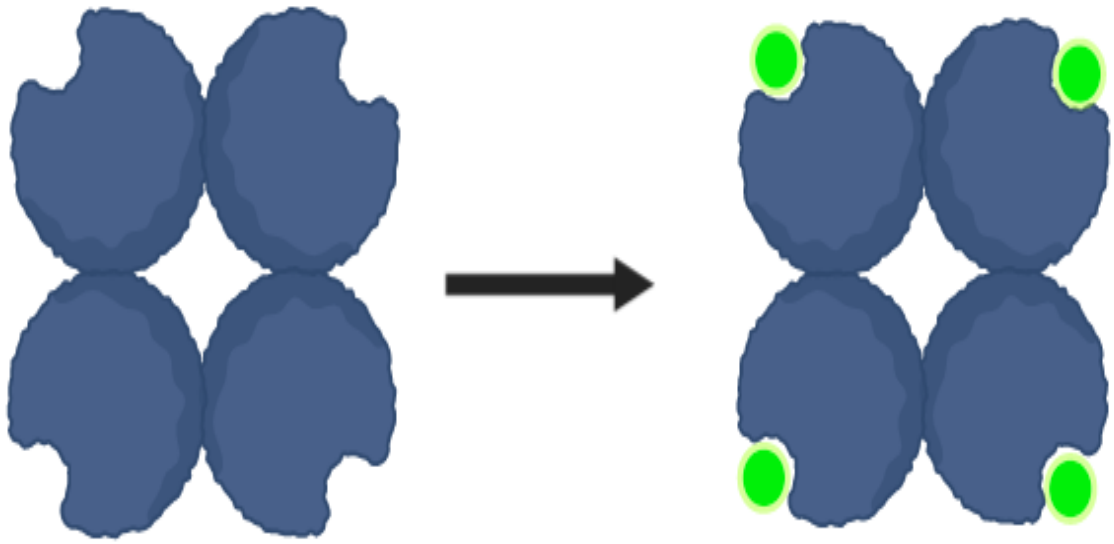

Figure 3.1: Schematic of the biotin-streptavidin interaction showing biotin (green circles) binding covalently to the streptavidin (blue tetramer). This is the strongest covalent interaction known in nature.

and issues with identifying *bona fide* cell-surface proteins in samples, rather than other intracellular membrane proteins [360] [361]. It is also possible that some cell surface biomarker proteins can be lost in the preparation of cells for M/S, as has been the case using multipotent human mesenchymal stromal cells (hMSCs) [360] . Neihage *et al.* (2011) [360] labelled the hMSCs with sulfo-NHS-SS-biotin, harvested the cells and purified them using streptavidin beads, thereby removing unlabelled cellular proteins and contaminants. The biotinylated proteins were eluted and the elution was run on a 1D gel before in gel trypsinisation and M/S. They identified 169 cell surface proteins that contained TMH or GPI anchors [360]. This is an almost identical approach to the one taken in this Chapter to examine *T. foetus* cells (Section 3.2.7). Biotinylation with a pull-down has also been used for parasites, such as *Toxoplasma gondii* [362]. However, there have been cases where avidin has bound non-specifically to the target cell wall proteins [363]. It has been thought that avidin may possess some structural similarity to cell-recognition domains and this non-specific binding may be indicative of cell-surface integrins [363].

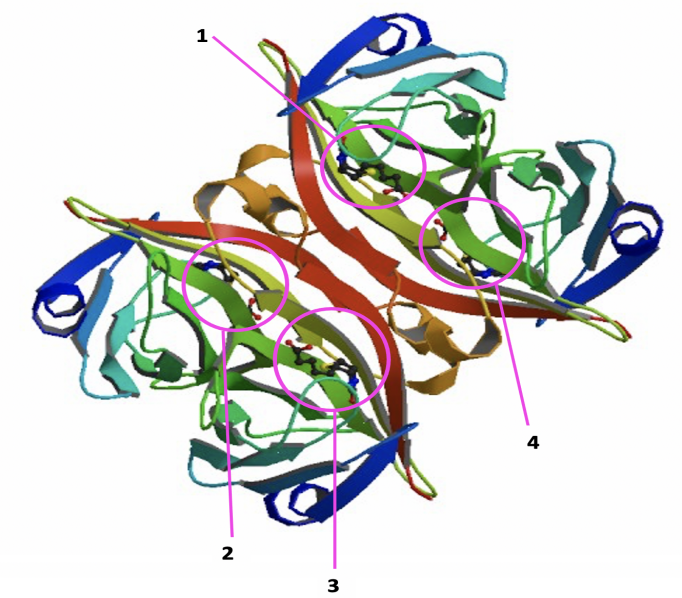

Figure 3.2: Structure of high affinity biotin binding to a streptavidin tetramer, the strongest covalent bond known. Each biotin molecule binds to one of the four streptavidin subunits. The four streptavidin molecules are highlighted with pink circles and numbered 1-4. Obtained from the Protein Data Bank (PDB) using the ID 1STP with a resolution of 2.6 Angstroms.

3.1.9 Peptide Arrays

Microarrays were originally created as a way to screen large amounts of DNA although [364] today they can be applied to a wide range of substrates including: peptides, microRNAs and tissues [365]. Peptide arrays produce a way to rapidly screen hundreds or thousand of peptides for various criteria such as immunogenicity, binding [305], or enzyme activity in one experiment [366] [367]. Peptide arrays can also be used to identify protein interactions [368], peptide-DNA interactions and peptide-cell interactions [369]. Microarrays are becoming increasingly widespread as a technique as the costs associated with the experiments fall.

The peptides are printed as spots in a known configuration onto a chip to which a secondary antibody conjugated to a reporter enzyme or fluorescent label is added. Once the substrate, for example, serum, is applied, and primary antibodies within the serum bind to their cognate peptides, a colormetric or fluorescent change occurs which can be imaged and quantified using specialist

software. Peptide arrays have been used for human diseases, such as to identify autoantigens in cancers [365] and also with regards to parasites, for example, serotyping *Toxoplasma gondii* [370]. Peptide arrays have also been used in attempts to identify malaria vaccine candidates [371], allowing information on the process of disease from various endemic regions to be gathered.

3.1.10 Aims and Objectives

The aim of this chapter is to validate the cell-surface location of *T. foetus* cell surface proteins predicted in Chapter 2 through mass spectrometry of cell fractions enriched for plasma membrane through biotin labelling, and through immunogenicity assays after exposure to sera from infected cattle. The chapter has three objectives:

1. Biotinylate the *T. foetus* cell surface, extract labelled proteins and identify cell-surface proteins using mass spectrometry

2. Compare the mass spectrometry profiles of trophozoites and pseudocysts to identify developmental differences in protein expression

3. Analyse the immunogenicity of selected cell surface proteins by screening serum from natural and experimental infections with a custom peptide microarray

3.2 Methods

3.2.1 *T. foetus* Cell Preparations for Label-free Mass Spectrometry

Three samples of Belfast strain *T. foetus* trophozoites and three samples of Belfast strain *T. foetus* pseudocysts all grown in Diamond media were prepared (approximately $1x10^7$ cells per sample). The samples were lysed using 200ml lysis buffer (1% SDS in 50mM ammonium bicarbonate) and the lysate was sonnicated three times for 10 seconds each time. The samples were then heated to 90°C on a heat block for 10 minutes and then centrifuged for 10 minutes at 13000rpm. The lysate was transferred to another tube, dilutions were made and a Bradford assay performed (Section 3.2.4).

3.2.2 Label-free Mass Spectrometry of *T. foetus*

Label-free mass spectrometry was carried out at the University of Liverpool Centre for Proteomic Research by Dr Stuart Armstrong. All samples were run into an SDS gel and then the band cut out and a semi-tryptic digest was performed [372]. Label-free mass spectrometry of the life stage samples was carried out by Stuart Armstrong using the Q -Exactive mass spectrometer and Orbitrap mass analyzer. Preferential expression of peptides was analysed by PEAKS 3 [373] [374].

3.2.3 SDS Page Gel Preparation

Unless stated otherwise, 12% SDS Page gels were used in all experiments. To produce two gels, the stacking gel comprised: 0.65ml Acrylamide, 1.25ml 0.5M Tris pH 6.8, 50μl 10% SDS, 3.025ml milliQ water, 25μl 10% ammonium persulphate, 2.5μl TEMED. The resolving gel was composed of: 4ml Acrylamide, 2.5ml 1.5M Tris pH 8.8, 50μl 10% SDS, 3.4ml milliQ water, 75μl 10% ammonium persulphate, 7.5μl TEMED.

3.2.4 Bradford Assay

The following concentrations of Bovine Serum Albumin (BSA) (μl/ml) were made up using milliQ water: 0,2.5,5,10,15,20 and 25 to produce the Bradford assay standards. A 96-well microplate was used for the assay. 100μl of the BSA standards were placed in duplicate down the left hand columns of the plate from lowest concentration to highest and 100μl the samples were placed in duplicate in the wells alongside the standards (Figure 3.3). 100μl of Bradford assay reagent was added to each

of the wells and left for 5 minutes at room temperature to incubate. The samples were then read and the concentrations calculated using a Tecan plate reader and Magellan software [375]. For both proteomic experiments, the trophozoite and pseudocyst samples were diluted: 1/200 and 1/400 and measured. The protein concentrations of the biotinylated soluble fractions were also diluted and measured using a Bradford Assay.

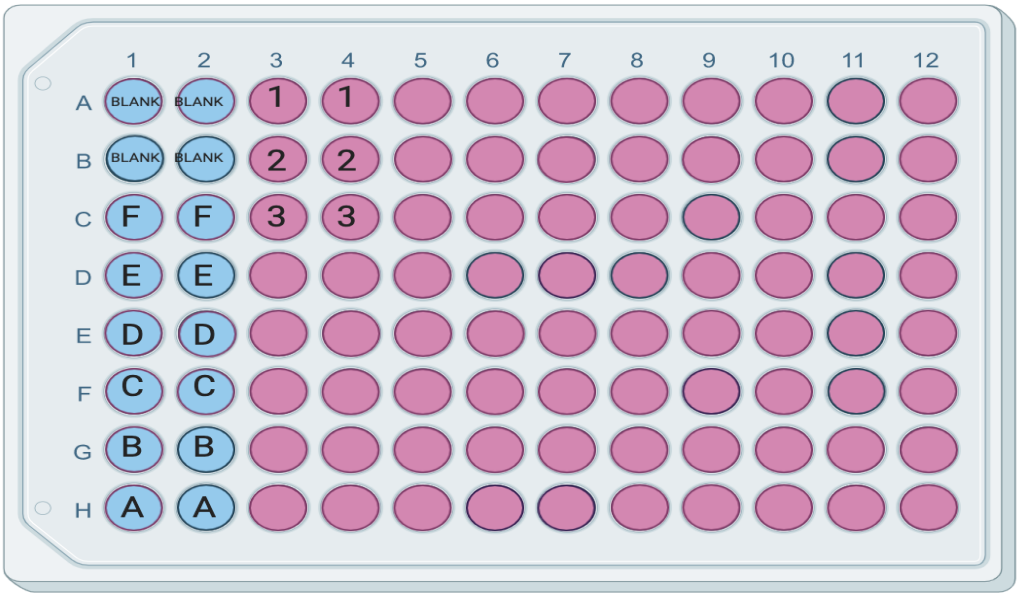

Figure 3.3: Schematic of the Bradford Assay 96-well plate layout showing the standards and blanks in blue and samples in duplicate in pink. Where blanks are milliQ water only, standard A is 25μl/ml BSA, B is 20 25μl/ml BSA, C is 15μl/ml BSA, D is 10μl/ml BSA, E is 5μl/ml BSA and F is 2.5μl/ml BSA. All samples are used in duplicate up to a maximum of 40.

3.2.5 Western Blots

The SDS page gels were run at 90 volts (V) for 2 hours. One gel was stained with Coomassie to show abundance of protein and the other was used in the Western blot.

The gel was washed in water for 5 minutes and then in transfer (Towbin) buffer (10x buffer = 30.3g/L Tris, 144g/L Glycine) for 5 minutes. A PDVF membrane was used and this was soaked in methanol for 10 minutes prior to use. The Western blot was run for 1 hour at 100V.

The membrane was washed in double distilled water (ddH20) and then stained with Ponceau S to check the transfer. The Ponceau S was then washed off and the membrane was covered with blocking buffer (5% milk protein in TBS) and left to incubate overnight at 4°C.

The membrane was removed from the blocking buffer and washed twice in Tris-buffered saline with Tween-20 (TBST) for 10 minutes. Streptavidin-HRP was made up in dilution buffer (1% milk protein in TBS) to a 1/5000 dilution and added to the membrane with rocking for 1 hour. The membrane was washed in TBST for 5 minutes twice, Tris buffered saline (TBS) for 5 minutes twice and ddH20 for 5 minutes twice. The membrane was dried and 4ml of Biorad clarity ECL substrate was added for 5 minutes. Images were taken with varying exposure times using the Biorad Chemidoc.

3.2.6 SDS Page Gel Staining

The SDS page gels produced were stained using both Coomassie blue (1.5 hours at room temperature with shaking. The excess stain was removing using three washes of 10 minutes of distilled water) and silver staining using the Pierce Silver Stain Kit from Thermofisher. The membrane was stained as per the manufacturers instructions with a final developing step of 2 minutes. The Western blot was stained using Ponceau S. The membrane was left in Ponceau S for approximately 3 minutes until bands were seen. The Ponceau S was then washed off using three washes of 2 minutes of distilled water and the bands were no longer visible before the membrane was incubated with the blocking buffer.

3.2.7 Biotinylation of *T. foetus* Cell Surface

Belfast strain *T. foetus* cells were grown to a density of approximately $1x10^7$ per ml in a 15ml tube. The cells were washed twice in PBS (Thermofisher) and spun down at 800xg at 4°C and re-suspended in PBS to give a total volume of cells of approximately $1.5x10^8$ in the 1ml volume. Cells were incubated with 0.5mg/ml EZ-link-sulfo-NHS biotin on ice for 45 minutes (Thermofisher A39257). Excess biotin quenched with 50mM Tris pH 7.5 and cells were washed three times with PBS, all wash steps using the centrifuge were performed at 4°C.
The biotinylation was also later performed using PBS at pH 7 and pH 9 to assess whether the acidicity of the solution effected the efficacy of the biotinylation.

3.2.8 Immunofluorescence Assay

100μl of the biotinylated cell solution was used after washing in PBS. The cells were fixed on 0.01% poly-L-lysin slides with 4% paraformaldehyde for 20 minutes and then the slides were rinsed with PBS twice. The cells fixed to the slides were blocked with PBS and 10% horse serum for 20 minutes at room temperature and then rinsed twice with PBS. The slides were incubated with streptavidin-FITC (1 in 100 dilution) in 10% horse serum in PBS for 1 hour in the dark at room temperature. The slides were then rinsed twice with PBS. Coverslips were mounted on the slides using Thermofisher prolong gold and 4',6-diamidino-2phenylindole (DAPI) then imaged on a Nikon Eclipse Tie fluorescence microscope using the Green Fluorescent Protein-2 (GFP2) and phase contrast channels.

3.2.9 Creation of Total Cell Extract

The cells were biotinylated as per the biotinylation protocol (Section 3.2.7). Cells were washed in PBS three times and resuspended in 1ml PBS. The resuspension was centrifuged for 5 minutes at 5,000xg at 4°C and resuspended in 100μl of ice-cold nuclease buffer (50μg/ml *Staphylococcus* nuclease, 2mM CaCl2, 20mM Tris-HCl pH 8.8, before use 1/1000 dilution of E-64 protein inhibitor was added). 10μl of 3% β-mercaptoethanol was added followed by 10μl DNase/RNase mix (1mg/ml DNase-1, 0.5mg/ml RNase-A, 50mM MgCl2, 0.5M Tris-HCl pH 7.0). The lysate was mixed with a 1ml syringe with a brown needle (26G). The lysate was frozen at -80°C and lyophilised before solublisation in 100μl sample buffer (160mM Tris-HCl, 25% sucrose,6.9% SDS, bromophenol blue).

3.2.10 Streptavidin Pulldown

The supernatant from the solubilised total cell extract (Section 3.2.9) was transferred to a 5ml tube. 4.5ml TENT-1% (50mM Tris-HCl, 5mM EDTA, 150mM NaCl, 1% Triton X-100) was added. The streptavidin-agarose beads were equilibriated with two washes in TENT-1%. 50μl per sample was used initially-then increased to 100μl per sample. The beads were added to the samples and incubated overnight with rotation at 4°C. The samples were spun down at 1000xg for 1 minute and both the beads and 1ml of the TENT-1%-cell suspension was transferred into a 1.5ml tube.

Initially the bead step was repeated-another 50μl of streptavidin-agarose beads was added to the remaining 3.5ml TENT-1%. The samples were incubated for 2 hours with rotation at 4°C, spun down and the beads and 1.5ml TENT-1% was transferred into a 1.5ml tube.

The beads were washed four times with 1ml TENT-1% and the proteins were eluted by heating at

95°C for 5 minutes with 100μl sample buffer in the presence of 10mM DTT. The beads were spun down for 5 minutes at 10,000xg and 10μl of each sample was then loaded onto SDS page gels.

3.2.11 *T. foetus* Cell Lysis

Initial attempts at proteomic analysis using the label-free mass spectrometry produced only a small number of peptides from the different *T. foetus* life stages, and so changes were made to the protocol. The first stage was to see whether cell lysis was responsible for poor yield. Three lysis buffers were made and tested on the *T. foetus* samples.

1. Lysis buffer 1- 50mM Hepes buffer at pH 8.0, 1% SDS, 1xcOmplete Protease Inhibitor Cocktail, EDTA (Sigma), 150mM NaCl, 10mM DTT
2. Lysis buffer 2 - 50mM Hepes buffer at pH 8.0, 8M urea, 1xcOmplete Protease Inhibitor Cocktail, EDTA (Sigma) 150mM NaCl, 10mM DTT
3. RIPA buffer - 50mM Tris, 150mM NaCl, 0.5% DOC, 1% NP-40, 0.1% SDS, pH 8.0 (In this case Triton-X-100 was used in place of NP-40)

For initial trials one tube (15ml) of *T. foetus* cells was grown for 24 hours to log phase to an approximate density of 5x10^5 per ml. Nine samples of 1ml were then taken and subjected to the different lysis buffers for different periods of time: 15 minutes, 30 minutes and 1 hour to see which condition was optimal for cell lysis and protein production. The tubes were spun down at 1000rpm for 5 minutes at 4°C. The media was removed and the cells were washed twice in PBS, each time the cells were spun at 1000rpm for 5 minutes. 500μl of lysis buffer was added to each pellet, three samples were given buffer 1, 3 buffer 2 and 3 RIPA buffer, and the pellet was re-suspended.
The tubes were then vortexed and then left for between 15 and 60 minutes with rotation at room temperature. Each buffer was left for each time point. The resulting samples were then spun at 1000rpm for 5 minutes and the pellet (insoluble fraction) and the supernatant (soluble fraction) were stored separately at -80°C.

It was determined that the RIPA buffer, left for 1 hour produced the highest amount of protein and so this was used for the main experiment. Ten samples, five for trophozoite and five for low temperature induced pseudocysts were used, 500μl RIPA buffer was used for 1 hour and the sample concentrations were measured using a Bradford assay.

3.2.12 TMT-labelled Mass Spectrometry Sample Preparation-Cell Lifestages

Cells were lysed using RIPA buffer and then sonicated. SP3 paramagnetic beads were used [376] as a clean-up strategy and to bind the proteins so any contaminants could be removed. The proteins were digested off the beads and then labelled with TMT. The samples were then fragmented by high pH and then reverse phase (R-P) chromatography [377] to produce 6 fractions. R-P chromatography used the hydrophobicity of peptides to get good separation. The samples were then run on the Orbitrap mass spectrometer.

3.2.13 Label-free Mass Spectrometry of *T. foetus* Biotinylated Samples

Samples of the eluted biotinylated proteins and the two controls were run into an SDS gel and then the band cut out and a semi-tryptic digest was performed as stated in Section 3.2.2. Label-free mass spectrometry was carried out as previously stated at the University of Liverpool Centre for Proteomic Research by Dr Stuart Armstrong. Preferential expression of peptides was analysed by PEAKS 3 [373] [374].

3.2.14 Mass Spectrometry Analysis

The mass spectrometry analysis of the peptides was performed using PEAKS [373], MaxQuant[378] and Perseus [379].

MaxQuant

The MaxQuant [378] analysis was performed using the following parameters: 1) No fractions; 2) Variable modifications of: Oxidisation of Methyonine and acetylation of N-terminus; 3) Fixed modifications- carbimidomethyl of C; 4) The maximum number of modifications per peptide was set at 5; 5) The machine used for the M/S was the Orbitrap, using all default parameters with a 4.5ppm tolerance; 6) the digestion method used is Trypsin/P with a maximum number of missed residues of 2; and 7) The peptides had to be greater than seven amino acids long with a max mass of 4600Da. All other parameters were set as default.

3.2.15 Peptide Chip Design

To examine the natural immunogenicity of putative *T. foetus* cell surface antigens, I designed a customized peptide microarray for screening antibody responses in bovine serum from natural

and experimental *T. foetus* infections. The microarray consisted of overlapping peptides from 51 proteins, each containing an N-terminal signal peptide (SP) in combination with a single C-terminal transmembrane helix (TMH), as well as asparagine (N) and threonine (T) residues that are predicted to be glycosylated, and all among the most abundant transcripts in trophozoite and pseudocyst cell culture. The array represented, therefore, the most abundant antigens predicted to be present on the parasite cell surface. The pipeline implemented to identify these proteins was similar to the process used in Chapter 2 to identify the cell surface proteome, as is shown in Figure 3.4.

There are 84,706 genes in the *T. foetus* genome sequence. These genes were translated and any shorter than 200 amino acids were removed to exclude dubious ORFS and mis-annotations. The remaining genes were analysed with SignalP 3 [380] and SignalP 5 [381]; 1219 were found to have a signal peptide. All 84,706 translated ORFs were also run through TMHMM [291] and 2,333 proteins had at least 1 TMH and 1179 had a single helix. Multi-spanning proteins were excluded because they make poor vaccine targets; the protein is embedded within, rather than projecting from, the membrane. Finally, the gene sequences were also run through PredGPI [293] to look for GPI anchors and 102 were found (with 99% reliability). A 'long list' of proteins comprising those with SP and a single TMH at the C-terminus were selected. The C-terminus position is necessary as the protein has to be pointing outwards from the cell surface, i.e. the N-terminus is outside. Using the TMHMM scores, only those TMH proteins with at most 10% of their total length inside the cell were retained; this excluding approximately 60% of TMHs.

These steps reduced the total number of eligible proteins from 1710 to 465, all of which were predicted to be type-1 transmembrane proteins, localised to the plasma membrane. The peptides were compared to the NCBI database using BLASTp [278] to detect any obvious homology that argued against cell surface roles. Any proteins that were obviously intracellular, for example, snare proteins and nucleoporins were removed based on this BLAST analysis. This reduced the list of 204 proteins.

It was anticipated that some of these proteins would belong to multi-copy gene families. The inclusion of multiple paralogues of a protein in the array would be inefficient, since they might be expected to give the same antibody response. To remove this potential source of redundancy, a BLAST database was created for the 204 proteins and then each was compared to the database individually using BLASTp. A threshold of e^{-25} was applied to remove close paralogues. In this way, no two spots on the array are very similar, so that the maximum structural diversity of protein can be screened.

After filtering by these steps, there still remained too many proteins to be accommodated on the chip, so the decision was made to reduce their number to 50 based on their abundance in cell culture. Highly expressed transcripts in culture are assumed to indicate highly expressed proteins in natural infections, and this is assumed to be beneficial for antibody recognition. Transcript abundance (FPKM) and variability thereof within experiments was determined from three low temperature and five control transcriptomes generated in Chapter 2. The mean and standard deviation of FPKM values were used to rank the protein sequences. The top 50 most abundant and least variable transcripts were then used for the peptide array.

3.2.16 Use of Transcriptomics for Peptide Array Design

Three low temperature and five control transcriptomes (generated in Chapter 1) were also used to identify peptides for the peptide array. The peptides in the array need to have a high abundance and have the smallest standard deviation across conditions. i.e. constitutively expressed genes across both life stages.

First, to trim the reads, Trimmomatic [382] on the Galaxy server [302] was used, using paired reads and all default settings. The reads were then mapped by Hisat2 [383] to the genome and GTF file. The resulting Hisat2 bam file was run through stringtie [384] using the unstranded setting. The isoforms were ordered and renamed and spreadsheet produced with the gene name and FPKM. The mean and standard deviation was added and the genes were ranked and cross-checked with the *in silico* selection. In this way the FPKM numbers from the transcriptomes were used to rank the genes by abundance and showed which were constitutively expressed across both the trophozoites and pseudocysts. The top 50 abundant and constitutively expressed genes were then used for the peptide array (Table 3.1).

Number	Gene ID	Product
1	TTF11197	hypothetical protein TRFO_10561
2	TTF35905	hypothetical protein TRFO_02143
3	TTF48034	hypothetical protein TRFO_40837
4	TTF39902	hypothetical protein TRFO_10466
5	TTF11874	hypothetical protein TRFO_05898
6	TTF55982	leucine-rich repeat domain-containing protein
7	TTF53402	hypothetical protein TRFO_12669
8	TTF44625	hypothetical protein TRFO_07982
9	TTF57869	hypothetical protein TRFO_29961
10	TTF19526	hypothetical protein TRFO_18803
11	TTF30114	hypothetical protein TRFO_22128
12	TTF07987	thioredoxin, putative
13	TTF36321	hypothetical protein TRFO_17724
14	TTF31573	hypothetical protein TRFO_02214
15	TTF56587	hypothetical protein TRFO_32199
16	TTF28278	hypothetical protein TRFO_35407
17	TTF07357	hypothetical protein TRFO_34912
18	TTF49921	hypothetical protein TRFO_12246
19	TTF03161	hypothetical protein TRFO_16017
20	TTF23499	Legume-like lectin family protein
21	TTF19157	cation channel sperm-associated protein subunit delta
22	TTF09721	adhesin-like protein
23	TTF00910	hypothetical protein TRFO_06137
24	TTF12012	hypothetical protein TRFO_39497
25	TTF14449	hypothetical protein TRFO_11445
26	TTF09063	hypothetical protein TRFO_39123
27	TTF44685	hypothetical protein TRFO_33209
28	TTF30148	Thioredoxin family protein
29	TTF72966	hypothetical protein TRFO_23368
30	TTF51333	hypothetical protein TRFO_33835
31	TTF41308	hypothetical protein TRFO_20972
32	TTF31692	hypothetical protein TRFO_38466
33	TTF14901	hypothetical protein TRFO_36551
34	TTF13877	hypothetical protein TRFO_19801
35	TTF38179	hypothetical protein TRFO_27265
36	TTF36584	leishmanolysin family protein
37	TTF03361	hypothetical protein TRFO_23519
38	TTF23900	Clan SC, family S28, unassigned serine peptidase
39	TTF73824	hypothetical protein TRFO_11378
40	TTF59933	Legume-like lectin family protein
41	TTF11679	hypothetical protein TRFO_27707
42	TTF75485	hypothetical protein TRFO_20015
43	TTF44949	hypothetical protein TRFO_32080
44	TTF36660	hypothetical protein TRFO_08766
45	TTF13670	hypothetical protein TRFO_19601
46	TTF34783	hypothetical protein TRFO_09395
47	TTF67688	hypothetical protein TRFO_40746
48	TTF16447	hypothetical protein TRFO_30251
49	TTF45090	hypothetical protein TRFO_35948
50	TTF75903	hypothetical protein TRFO_26675

Table 3.1: *T. foetus* proteins included in the peptide array

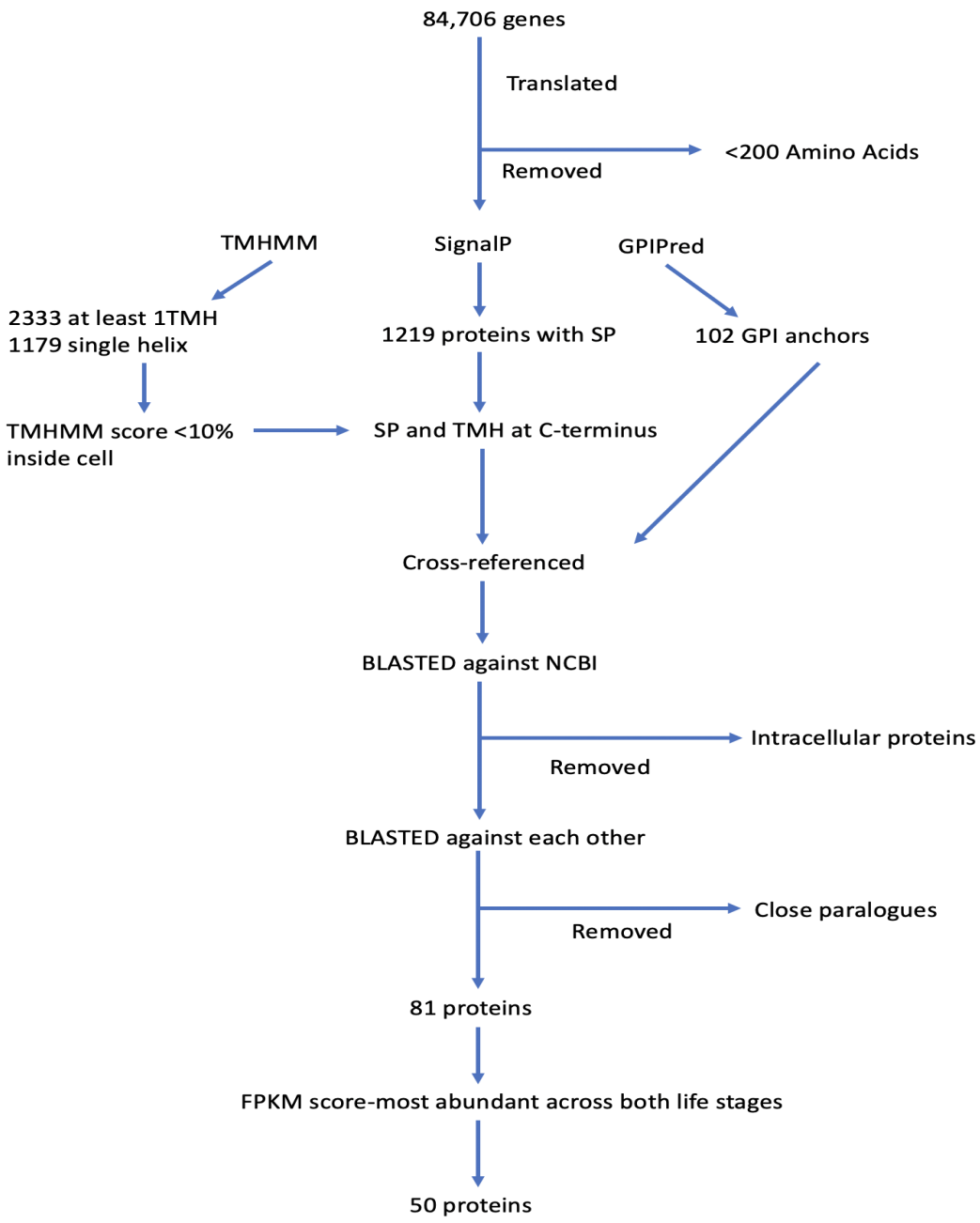

Figure 3.4: Flow diagram of the peptide array design from the full genome of 84,706 ORFs to a set of 50 proteins. These proteins were over 200 amino acids in length with a signal peptide and one transmembrane domain, both at the C-terminus of the peptide. They were not close paralogues of each other nor intracellular proteins.

3.2.17 Peptide Array

The peptide microarray was manufactured by PEPperPRINT [385]. Each of 50 proteins was split into peptides 15 amino acids in length with an overlap of 13 amino acids. In total, each array contains 5,321 peptides printed in duplicate to give 10,642 peptides in total across the chip. Control peptides allow for correct function to be confirmed in the assay; the hemagglutinin A (HA) control was YPYDVPDYAG and was used on 116 spots and the polio control was KEVPALTAVETGAT and used on 114 spots per chip giving a total of 230 control peptides. The peptides were linked and elongated with neutral GSGSGSG linkers at the C- and N- terminus and translated into overlapping peptides. The peptides were distributed across three panels on each of six chips (PEPperPRINT slides array codes: 002492_02 to 002492_07)

Reagents

The four buffers used were:
1) Standard buffer: PBS, pH 7.4, 0.05% Tween-20, filtered
2) Blocking buffer: Standard buffer with the addition of 1% BSA, filtered
3) Staining buffer: Standard buffer with the addition of 10% blocking buffer, filtered
4) Dipping buffer: 1mM Tris, pH 7.4

Conjugate:
Goat anti-bovine IgG (H+L) Cy3 (Jackson ImmunoResearch Laboratories)
Control conjugate (provided by PEPperPRINT): Mouse monoclonal anti-HA (YPYDVPDYAG) Cy5 Sera from naturally and experimentally infected cows were used for primary antibodies, these cows had previously been diagnosed as *T. foetus* positive. Pre-immunization sera from 5 experimentally infected bovines were used as negative controls. Sera collected post immunization from the same animals (n=5) were used as positive samples. In addition, sera from 10 naturally infected bovines were used as positive samples. All sera were combined in pools (Table 3.2) and incubated on a different chip, two replicates of each. The sera samples were provided by Dr Matt Brewer of the University of Iowa. The sera from the naturally infected animals was opportunistically collected when a naturally positive animal was found. The experimental calves were challenged with a field strain isolated from a bull in Iowa, USA and serum was collected approximately 30 days post infection.

1) Pool of pre-immune sera (n=5)

2) Pool of naturally infected cattle (n=5)

3) Pool of experimentally infected cattle (n=10)

3.2.18 Immunoassay

Each peptide microarray slide was placed into an incubation tray. 1ml of standard buffer (PBS, pH 7.4, 0.05% Tween-20, filtered) was added to the corner of each microarray chamber and incubated for 15 minutes at room temperature on an orbital shaker (140rpm). The standard buffer was then aspirated from the corner of the chamber. 1ml of blocking buffer (standard buffer with the addition of 1% BSA, filtered) was then added to each microarray via the corner of the chamber. A lid was applied to the tray that included a waterproof seal. The slide was incubated at room temperature for 45 minutes on an orbital shaker (140rpm). The liquid was then aspirated from the corner of the array.

One slide was pre-stained with a secondary antibody. Goat anti-bovine IgG (H+L) Cy3 (Jackson ImmunoResearch Laboratories) was diluted in staining buffer 1:4,500 and 1ml was added to the microarray. The lid was replaced. The microarray was incubated for 30 minutes at room temperature or an orbital shaker (140rpm) in the dark, after which the liquid was aspirated from the chamber. The array was washed three times for 1 minute each time with 1ml of standard buffer on the orbital shaker (140rpm). The slide was removed from the tray and dipped twice in dipping buffer then dried in a stream of air. The microarray was then analysed in a microarray scanner (Section 3.2.19). This was in order to acquire a background reading of the ambient levels which could later be used in a background correction step to obtain the true spot intensities.

Each slide was treated with bovine serum samples, as described in Table 3.2. The negative control was sera from 5 cattle that had not been exposed to *T. foetus* (pre-immune). The same 5 cows were later experimentally infected with *T. foetus* and sera was collected to use as a positive sample (post-immune sera) and sera from 10 naturally infected cows was also used as another positive sample. There were two replicates of each sample which were to later be averaged. The sera was added to both the new slides that were not incubated in the pre-staining step and also the slide that was used for the pre-staining step. The new slides must be incubated with the standard buffer and blocking buffer as per the pre-staining protocol to hydrate and block the slides before use.

To equilibrate the slide, 1ml of staining buffer was added to the microarray and the lid was applied. The microarray was incubated for 15 minutes at room temperature on an orbital shaker (140rpm).

The liquid was then aspirated from the corner of the chamber. The microarray was incubated at 2-8°C overnight on an orbital shaker (140rpm). The liquid was then aspirated from the corner of the chamber. The array was washed three times for 1 minute each time with 1ml of standard buffer on the orbital shaker (140rpm). Both the secondary antibody goat anti-bovine (1:4500 dilution) and control antibody anti-HA (1:1000 dilution) were mixed in staining buffer. 1ml was added to the microarray and covered by the lid. The microarray was incubated for 60 minutes at room temperature on an orbital shaker (140rpm). The liquid was then aspirated from the corner of the chamber. The array was washed three times for 1 minute each time with 1ml of standard buffer on the orbital shaker (140rpm). The slide was removed from the tray and dipped twice in dipping buffer then dried in a stream of air. The slide was then examined in a microarray scanner (Section 3.2.19).

Chip	Chip 1	Chip 2	Chip 3	Chip 4	Chip 5	Chip 6
Array Code	002492_02	002492_03	002492_04	002492_05	002492_06	002492_07
Replicate	1	2	1	2	1	2
Sample	Pool of pre-immune sera from 5 experimentally infected	Pool of pre-immune sera from 5 experimentally infected	Pool of post immune sera from 5 experimentally infected	Pool of post immune sera from 5 experimentally infected	Pool of sera from 10 naturally infected	Pool of sera from 10 naturally infected
Sample Codes	Heifers: 70, 3263, 3265, 3266, 3275	Heifers: 70, 3263, 3265, 3266, 3275	Heifers: 70, 3263, 3265, 3266, 3275	Heifers: 70, 3263, 3265, 3266, 3275	Cow1-Cow10	Cow1-Cow10
Positive/ Negative	Negative	Negative	Positive	Positive	Positive	Positive

Table 3.2: List of samples used for the detection of significant epitopes with the peptide array immunoassay

3.2.19 Peptide Array Imaging and Quantification

An Agilent G2565CA Microarray Scanner (Agilent Technologies, USA) was used at the University of Liverpool, to record fluorescence emitted by the labelled slides. For both the pre-staining step and the staining with the secondary antibody, the slides were analysed for red (670nm) and green (570nm) channels independently with 10μm resolution. Images were saved with 16-bit grayscale as tiff files (1 image per channel). The tiff images (both red and green) produced were opened in the PEPSlide Analyzer software (Sicasys Software GmbH, Heidelberg, Germany). The program overlays both images to allow the user to see both colours simultaneously. The images were rotated

and filtered for noise to allow the spots to become more visible.

Accurate quantification of fluorescence requires that the image be aligned with a map file containing the coordinates and identity of each spot. The array map was aligned with the image using the control spots as a guide. The background correction level 'global' was used [386]. This was to have a global background for the entire array based on the background pixels of all spots in the array. For each spot in the microarray, the mean and median of its raw, background and foreground fluorescence values in both the red and green channels were calculated. The data was saved as a .gpr file which is needed for the limma analysis (Section 3.2.20). 18 files were produced in total as there were two slides, each with three microarrays and the data quantification was performed three times.

3.2.20 Peptide Array Analysis

Limma [387] in R [308] was used to create a linear model [388] [389] . This was is order to model the relationship between the spot intensities of negative and positive samples. The source file used was from the 'genepix' analysis software [389] and a one channel analysis, using only the green fluorescence channel, was used, meaning, only the green spots were read and used in the analysis rather than both red and green. The control spots were read by limma but were removed later in the process and so were not included in the analysis. The analysis was performed three times, once per microarray map. For background correction the 'normexp' method was selected as this is optimal for using .gpr files. The background correction is used to remove the background intensity, i.e. the ambient signal, such as from non-specific binding, from the foreground spot intensity [387] [386]. In this way it prevents negative or zero values which would otherwise affect the linear model. Normalisation [389] was performed using variance stabilising normalisation (VSN) [390] in the limma package to transform the data. Peptide filtering was performed to remove the control spots and the intensity from the duplicate spots were averaged. To fit the linear model, a design matrix was made to compare the pre-immune sera (negative control) with the naturally and experimentally infected samples (positive). The standard errors from the log-fold change were normalised using the empirical Bayes method [391] [392]. This was to reduce differences between replicates due to technical differences rather than biological ones.

To identify significant immunogenic epitopes an adjusted p-value<0.05 (padj)was used. The false discovery rate (FDR) was estimated using the Benjamini-Hochberg method (log2fold-change > 2). The peptides were ranked based on their padj values as well as their fold change (FC), generating a

list of the upregulated and downregulated peptides. The same data analysis methods were carried out on the other two microarray maps.

3.3 Results

3.3.1 Label Free Mass Spectrometry of Lifestages

Both *T. foetus* trophozoites and low temperature induced pseudocysts were analysed using label-free mass spectrometry to identify differentially expressed proteins between life-stages. In total 323 proteins were identified using label-free mass spectrometry, that had a fold change over 2 and at least 2 unique peptides associated with them. In total 316 proteins had higher expression in the pseudocyst sample relative to the trophozoite sample and 7 had higher expression in the trophozoite sample relative to the pseudocyst sample (Table 3.3). Two results are domain of unknown function (DUF) (TTF14018 and TTF29497). Ornithine Transcarbamylase (TTF32993) is found in the biosynthesis pathway for arginine. The Iron Superoxide Dismutase (TTF58332) may be involved in the response of *T. foetus* to environmental oxygen [297] and also may be facilitated by iron, which is known to be a key element required by *T. foetus* [263].

Gene ID	Product	Fold Change
TTF58332	Iron Superoxide Dismutase	5.49
TTF32993	Ornithine Transcarbamylase	2.82
TTF14018	DUF1900-domain-containing Protein	2.21
TTF29497	DUF4980 domain-containing Protein	2.19
TTF26374	WD40 repeat-like Protein	2.16
TTF57322	Hypothetical Protein TRFO_0662	2.06
TTF19653	Cysteine Protease 3	2.02

Table 3.3: Fold change of *T. foetus*trophozoites relative to pseudocysts from label-free mass spectrometry experiments. Pseudocysts were induced by the lowering of the environmental temperature to 4°C. Samples were run on the Orbitrap mass spectrometer and data was first analysed using Peaks [373] and then exported to Progensis [374]. Results included had a fold change over 2 and at least 2 unique peptides associated with each.

Of the top 30 pseudocyst proteins with the highest fold change (Table 3.4) compared to the trophozoites, 8 are hypothetical proteins and one (TTF15093) has a product of 'NA' meaning that it does not have a structural similarity to any protein currently in the NCBI database (Table 3.3). There are two clan CA proteins (TTF58150 and TTF16060) which would be expected since this is a very large gene family in the *T. foetus* genome. There are also two adhesin-like proteins (TTF82447 and TTF05494) but no clear proteins that would suggest a difference in morphology or metabolism.

The fact that so few proteins were identified as having higher expression in the control trophozoites compared to the pseudocyst makes comparisons between the two difficult. There does not seem to be any clear differences in terms of protein families or functions.

Gene ID	Product	Fold Change
TTF62432	WH2 Motif Domain Containing Protein	315.81
TTF63297	RNA-binding Protein	60.25
TTF82447	Adhesin-like-protein	60.15
TTF23865	Geranylgeranyl transferase type-2 subunit beta-like	50.42
TTF60517	Hypothetical protein TRFO_29767	43.33
TTF48222	Synaxin-8	41.37
TTF59123	Hydrogenase, Fe-only	32.35
TTF51390	Mitotic apparatus protein	30.04
TTF16060	Clan CA, family C1, cathepsin L-like cysteine peptidase	27.96
TTF47485	Succinyl-CoA ligase [GDP-forming] subunit alpha, mitochondrial	27.57
TTF21656	Co-chaperone GroES	24.40
TTF23767	Hypothetical protein TRFO_17212	23.19
TTF13778	Hypothetical protein TRFO_33979	22.09
TTF62170	Small GTP-binding protein	19.48
TTF36735	dnaK protein	18.23
TTF29274	Hypothetical protein TRFO_21306	18.21
TTF41140	60S acidic ribosomal protein P1	17.26
TTF05494	Adhesin-like protein	16.06
TTF81870	Hypothetical protein TRFO_14461	15.38
TTF08565	Hypothetical protein TRFO_20073	15.20
TTF82812	EF hand family protein	15.12
TTF67499	60S acidic ribosomal protein P1	15.03
TTF58150	Clan CA, family C1, cathepsin L-like cysteine peptidase	14.47
TTF00205	T-complex protein 1 subunit alpha	14.14
TTF00474	Eukaryotic translation initiation factor 2 subunit alpha	14.07
TTF15822	Activator of Hsp90 ATPase, putative	13.87
TTF75555	Hypothetical protein TRFO_37932	13.55
TTF66092	Hypothetical protein TRFO_22858	13.33
TTF15093	-NA-	12.95
TTF11891	HMG box family protein	12.74

Table 3.4: Top 30 *T. foetus* pseudocyst proteins showing the highest fold change relative to trophozoites from label-free mass spectrometry experiments. Pseudocysts were induced by the lowering of the environmental temperature to 4°C. Samples were run on the Orbitrap mass spectrometer and data was first analysed using Peaks [373] and then exported to Progensis [374]. Results included had a fold change over 2 and at least 2 unique peptides associated with each.

3.3.2 TMT Mass Spectrometry of Lifestages

Due to the low numbers of identified proteins using the label-free approach, another attempt was made to identify preferentially expressed proteins between the lifestages, using TMT labelled mass spectrometry.

There are very few differences between trophozoite and pseudocyst peptides, hypothetical proteins

were commonly found to be expressed in both samples. This is not surprising since there are 12,114 'hypothetical' genes in the genome. There did seem to be more calcium related genes in the pseudocysts and a limited number of heat shock proteins. However only 138 peptides were found to have significant differences in expression (proteins significance 20, false discovery rate of 1% and ratio corresponding to a minimum fold change of 1.5) between all samples which is much lower than expected (Table B1). There is also a large number of protein kinases found to be upregulated in both the trophozoite and pseudocyst samples, which are known to be a large *T. foetus* gene family, and akyrin containing proteins. Ratios were calculated relative to the trophozoite sample intensities, this corresponds with fold change. In total 53 proteins had a higher expression in the pseudocyst than trophozoite and 85 had higher expression in the trophozoite than the pseudocyst.

3.3.3 Cell Surface Biotinylation and Streptavidin-pulldowns

The cell surface of *T. foetus* trophozoites and pseudocysts were labelled using biotin and then extracted using a streptavidin pulldown. This was in order to produce a sample of cell surface expressed proteins only that could then be analysed using mass spectrometry. When imaged, the biotinylated cells showed clear fluorescence but the non labelled cells did not (Figures 3.5 and 3.6) showing that the cell-surface labelling was successful. Furthermore, the Western blots also show that the cells have been biotinylated (Figure 3.7 and 3.8) as bands are present in the biotinylated lanes. In Figure 3.9 there are no bands in the control lanes but there are bands in all of the elutions from the streptavidin-pulldowns showing that the samples have been successfully labelled and biotinylated proteins have been successfully recovered. Figures 3.10 and 3.11 also show that this pattern on the Western blot is due to biotinylation and not purely protein quantity as the SDS page gels have both been stained and show clear banding for all samples, confirming there is an amount of protein present in each.

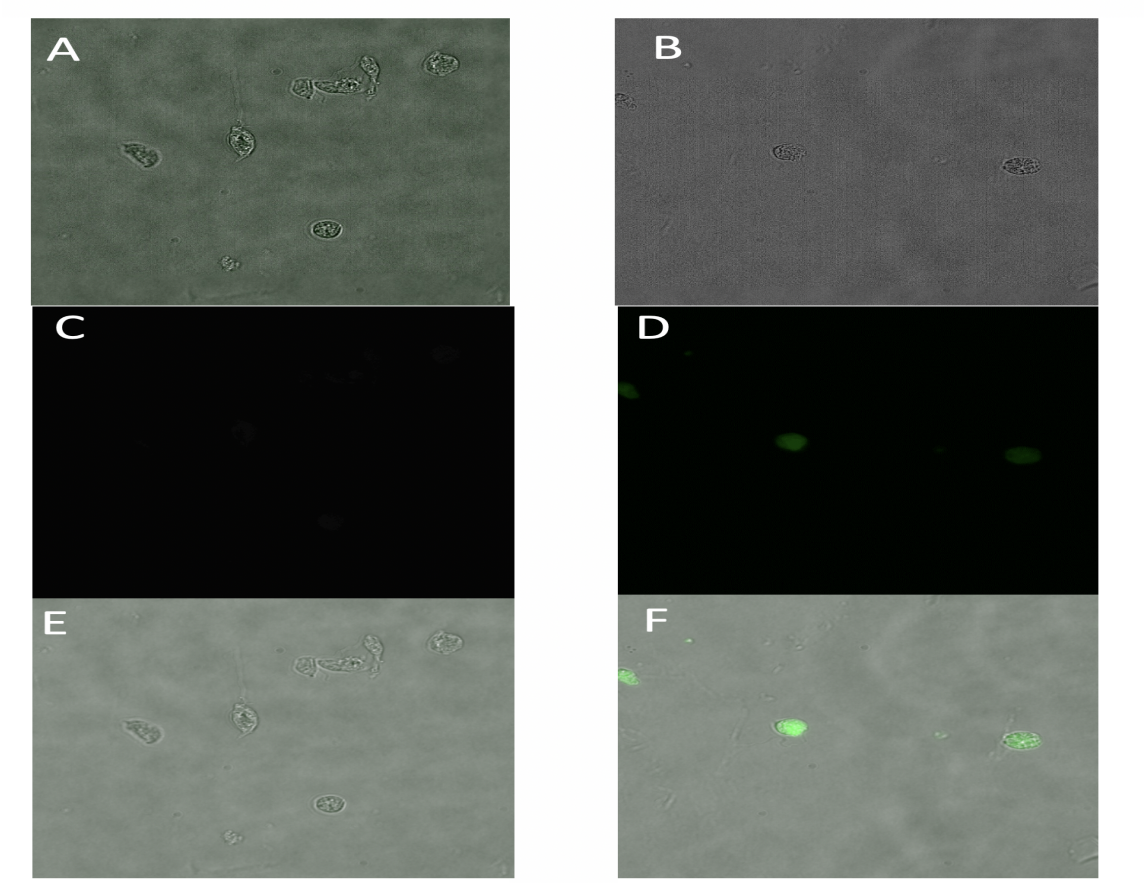

Figure 3.5: Biotin Labelling of the *T. foetus* cell surface. *T. foetus* trophozoites, both control (unlabelled- panels A,C and E) and labelled (panels B, D and F) after 30 minutes of labelling with 0.5mg/ml sulfo-biotin-NHS. After biotinylation the cells were incubated with 1/100 dilution of 1mg/ml streptavidin-FITC and mounted onto poly-L-lysine coated slides, fixed with 4% paraformaldehyde and mounted with Prolong gold. Images were taken on a Nikon eclipse Tie microscope using the Phase contrast (A and B) and GFP2 (C and D) channels. A composite image was also produced using both channels (E and F) showing that the labelled cells are fluorescing whereas the control cells are not.

3.3.4 Comparison of Biotinylation of Trophozoites and Cysts

There did not seem to be any marked differences between the trophozoite and pseudocyst western blot results (Figure 3.7, 3.8, 3.9)and both life stages showed fluorescence. When elutions from all three trials were compared, there were again clear bands on the biotinylated samples and not

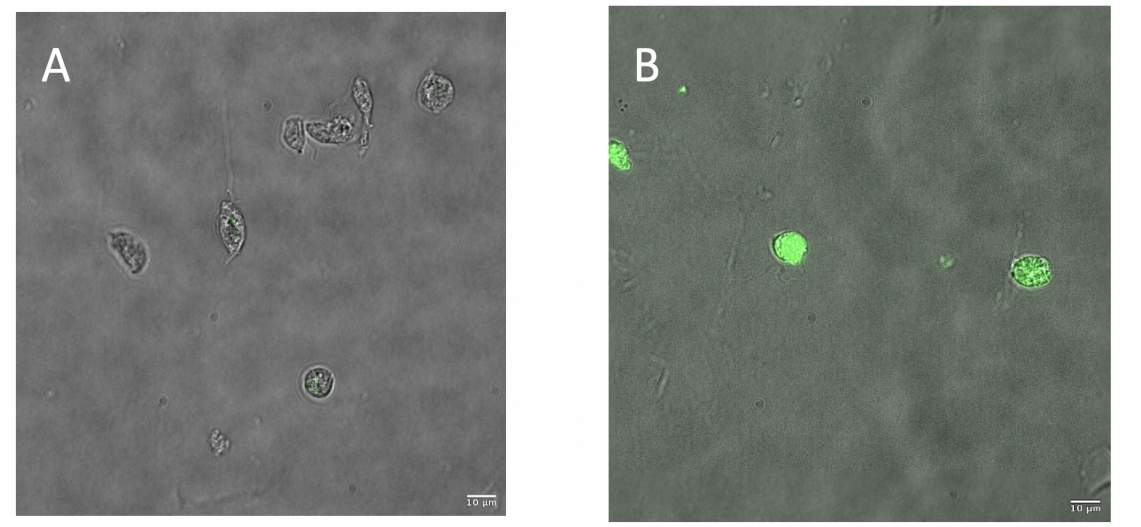

Figure 3.6: Biotin labelled and unlabelled *T. foetus* cells. Images were taken on a Nikon eclipse Tie microscope using the Phase contrast and GFP2 channels. A composite image was also produced using both channels. The composite images from Figure 3.5 (E and F) were corrected for sharpness and contrast in ImageJ.

on the control (Figure 3.9) but there was also clear banding on the SDS page gel labelled by both Coomassie (Figure 3.10) and Silver staining (Figure 3.11) and showing that there was protein loaded on all samples. When samples were not biotinylated there were no bands seen on the Western blots.

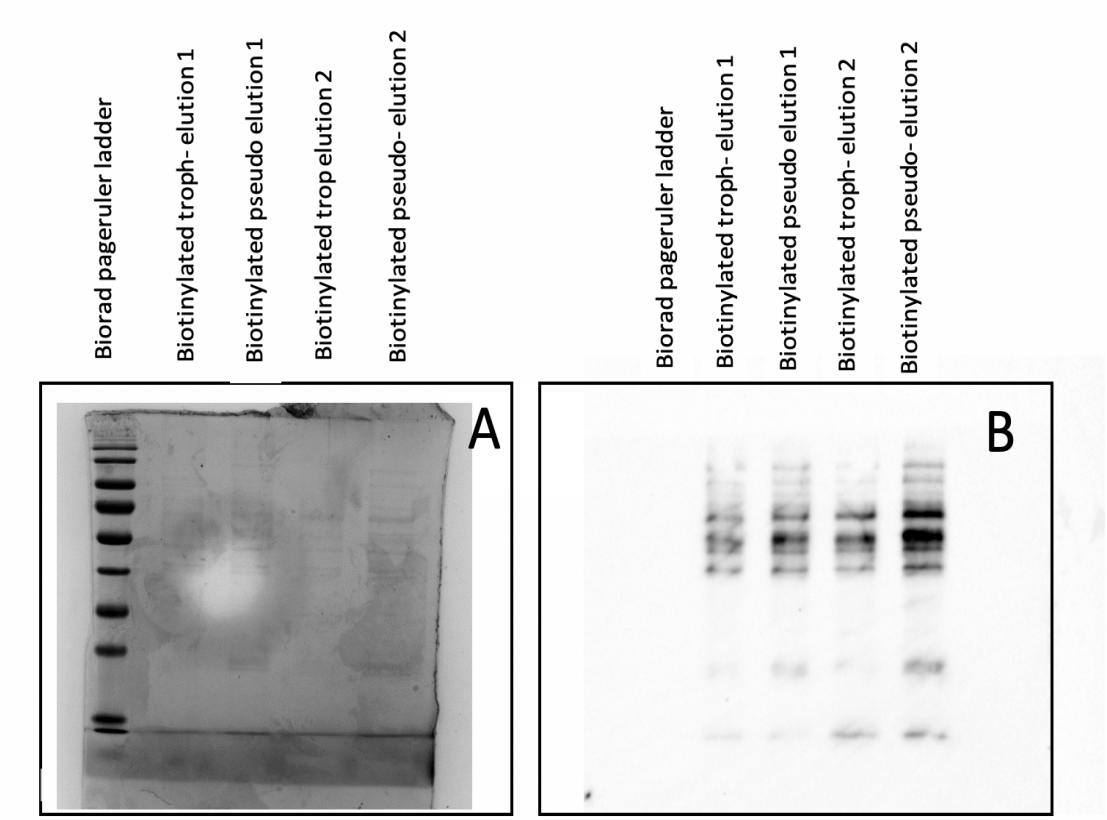

Figure 3.7: A: Coomassie stain of biotinylated *T. foetus* samples.
B: Western blot of biotinylated *T. foetus* samples.
For both trophozoites (trophs) and pseudocysts(cysts) approximately $1x10^8$ cells were used each time. Samples were labelled using EZ-link-NHS-biotin for 45 minutes at 4°C. They were then subjected to a nuclear protein exclusion protocol and a streptavidin pulldown was performed using streptavidin-agarose beads. The proteins were eluted off the beads by boiling in 100μl sample buffer in the presence of DTT for 5 minutes. The elution was stored and the beads were eluted from again using the same process. This gave two samples of 100μl of elution (elution 1 and elution 2) for each condition, both trophozoites and pseudocysts. 10μl of sample was then run on a 12% SDS page gel and a Coomassie stain and a Western blot was performed. For the Western blot, the membrane was imaged on the Chemidoc with an exposure time of 54 seconds.

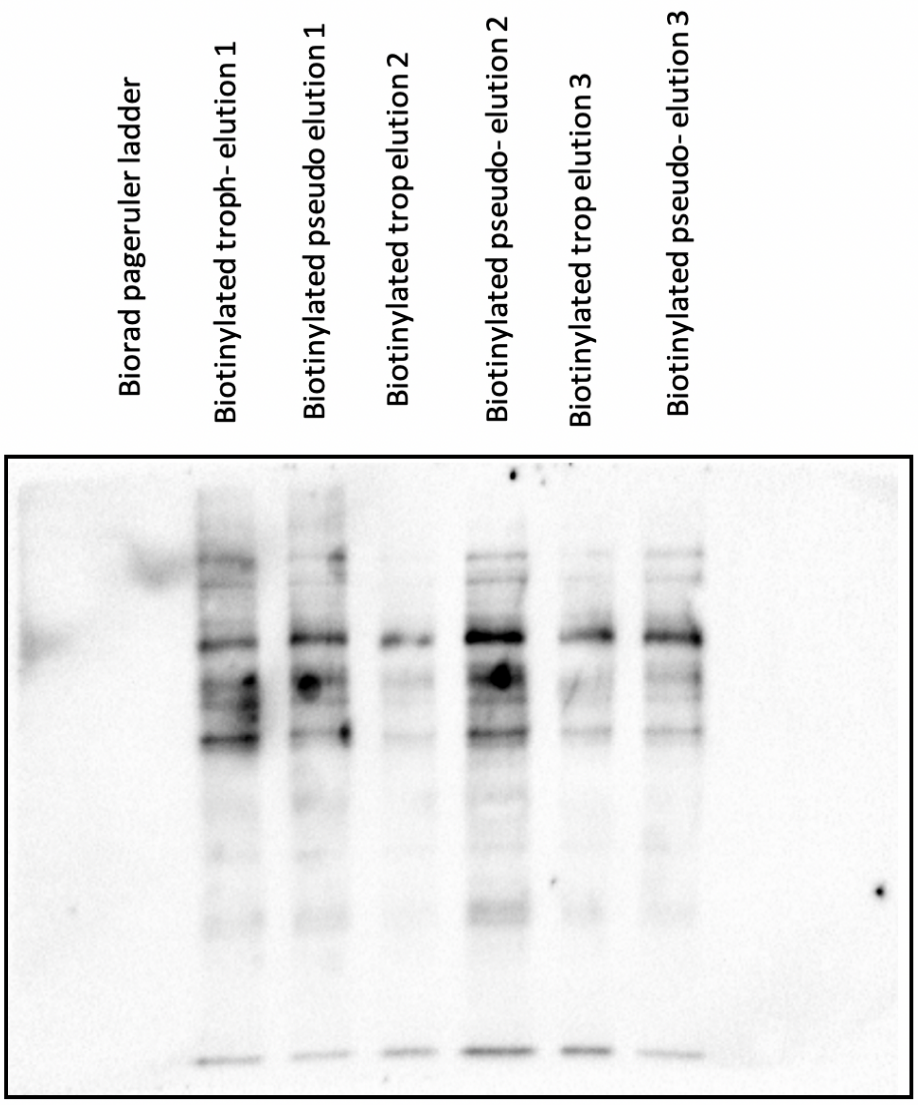

Figure 3.8: Western blot of biotinylated *T. foetus* trophozoites and pseudocysts. Both trophozoites (trophs) and pseudocysts(cysts) were used and approximately 1x10$\hat{8}$ cells were used each time. Samples were labelled using EZ-link-NHS-biotin for 45 minutes at 4°C. They were then subjected to a nuclear protein exclusion protocol and a streptavidin pulldown was performed using streptavidin-agarose beads. The proteins were eluted off the beads by boiling in 100µl sample buffer in the presence of DTT for 5 minutes. The elution was stored, and the beads were then eluted from again using the same process. This was repeated again to give 3 samples of 100µl of elutions (elution 1, elution 2 and elution 3) for each condition, both trophozoites and cysts. 10μl of sample was then run on a 12% SDS page gel and a western blot was performed. For the western blot, the membrane was imaged on the chemidoc with an exposure time of 10 seconds.

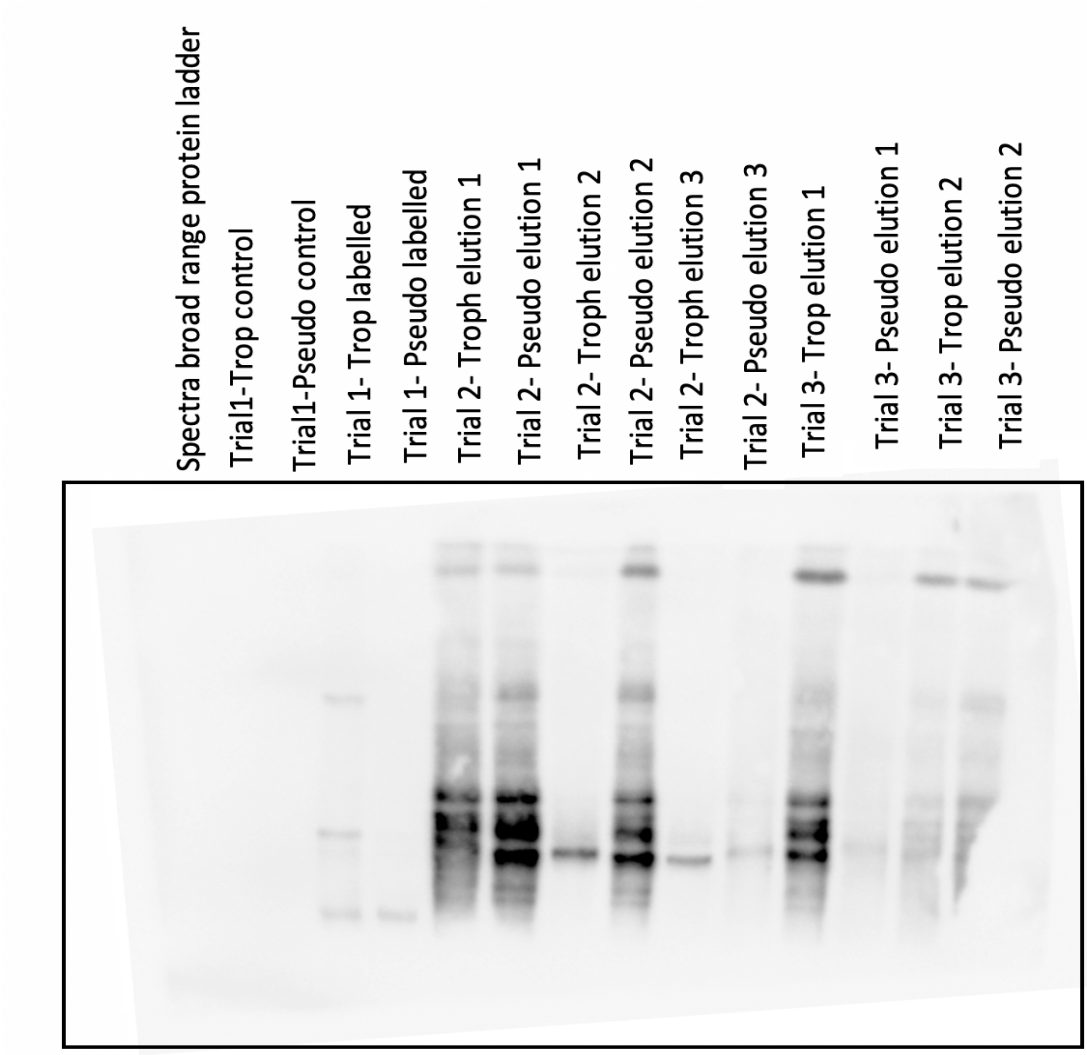

Figure 3.9: Western blot of three biotinylation trials of *T. foetus* trophozoites and pseudocysts. For each sample the cells were centrifuged and the nuclear fraction was removed. Lane labelled 'Trial' contain the samples in which the cells have been biotinylated and the proteins 'pulled down' by way of a streptavidin pulldown using streptavidin agarose beads. The lanes labelled 'control' contain unlabelled *T. foetus* cells that were processed in the same way as the labelled samples.

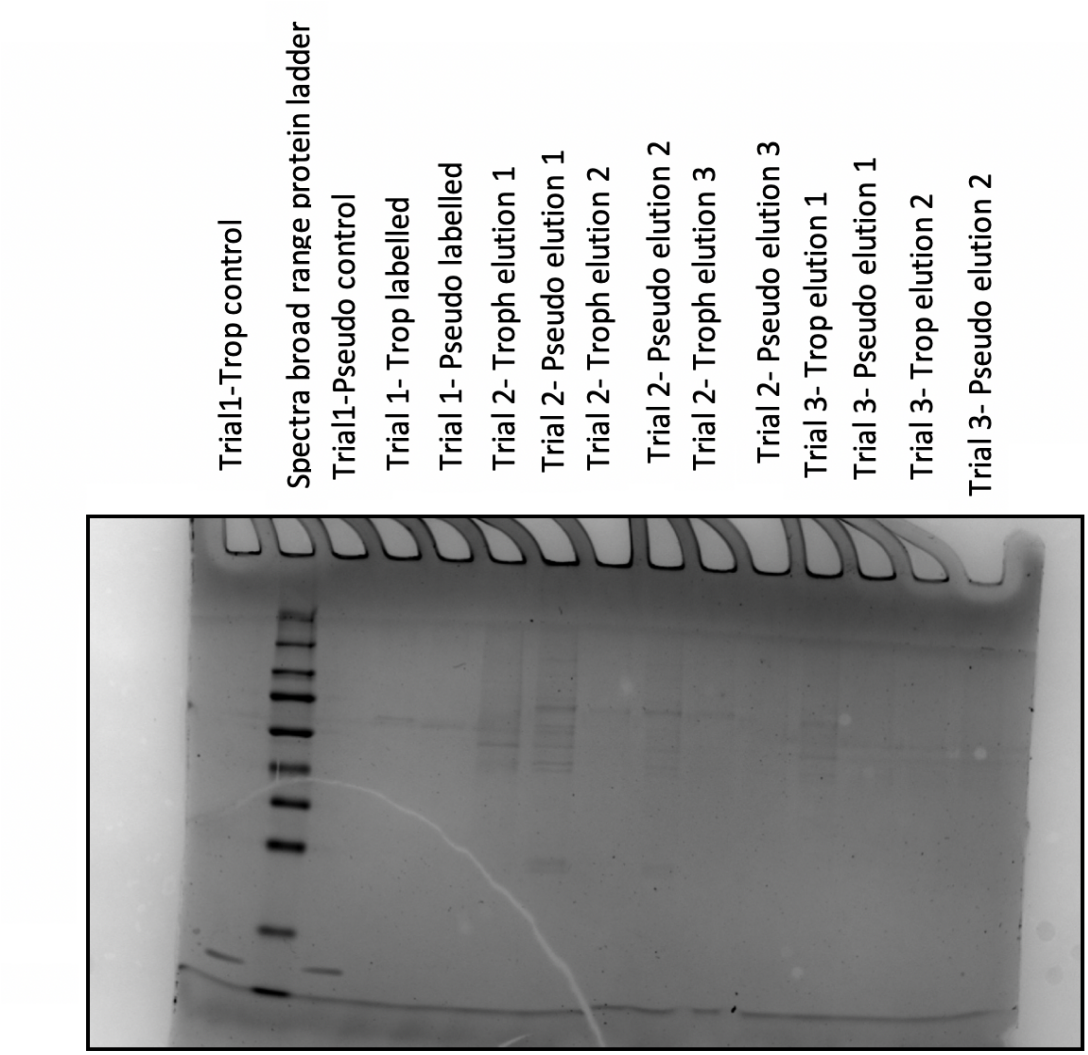

Figure 3.10: SDS page gel of three biotinylation trials of *T. foetus* trophozoites and pseudocysts. A Coomassie stain was used to show the protein bands. For each sample the cells were centrifuged and the nuclear fraction was removed. Lanes labelled 'Trial' contain the samples in which the cells have been biotinylated and the proteins 'pulled down' by way of a streptavidin pulldown using streptavidin agarose beads. The lanes labelled 'control' contain unlabelled *T. foetus* cells that were processed in the same way as the labelled samples.

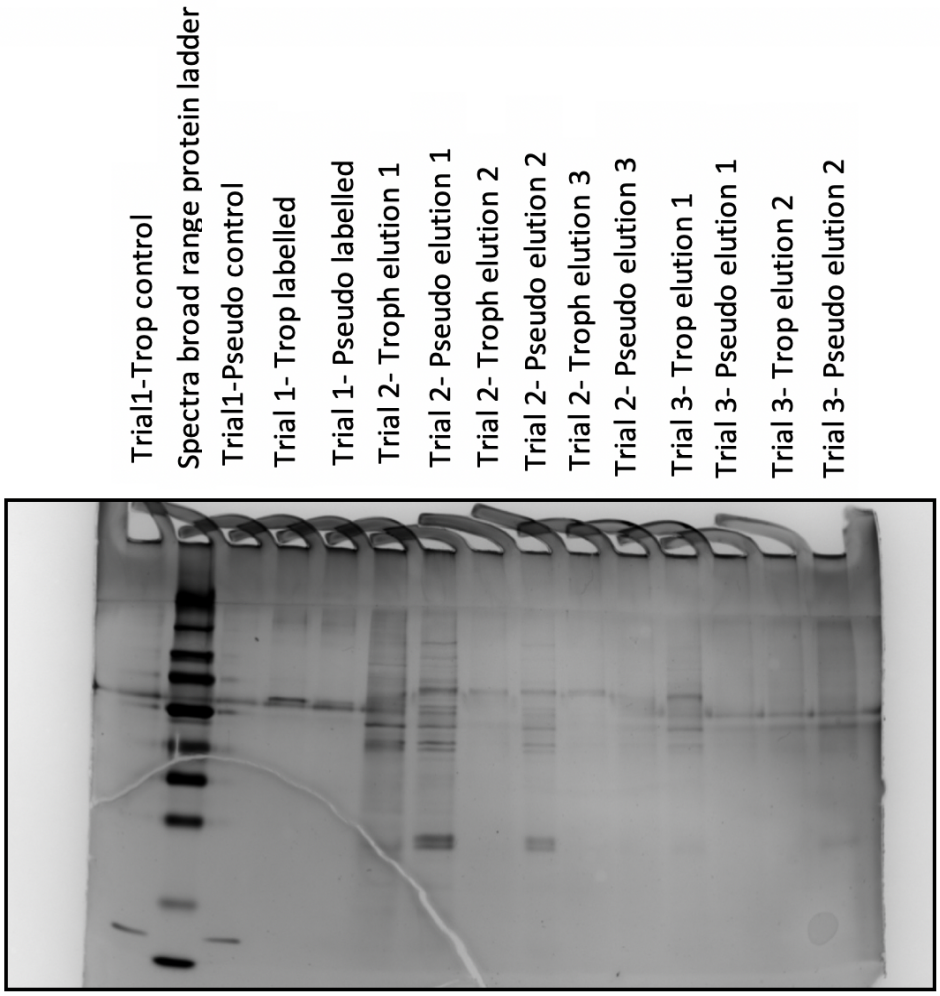

Figure 3.11: SDS page gel of three biotinylation trials of *T. foetus* trophozoites and pseudocysts. A silver stain was used to show the protein bands. For each sample the cells were centrifuged and the nuclear fraction was removed. Lanes labelled 'Trial' contain the samples in which the cells have been biotinylated and the proteins 'pulled down' by way of a streptavidin pulldown using streptavidin agarose beads. The lanes labelled 'control' contain unlabelled *T. foetus* cells that were processed in the same way as the labelled samples.

3.3.5 Label-free Mass Spectrometry of Biotinylated Proteins

When the *T. foetus* cells were biotinylated, three trials were performed for each lifestage and both trial 2 and trial 3 had multiple elutions for each lifestage. All elutions along with unlabelled control samples were analysed using label-free mass spectrometry. This was to identify which proteins appear to be found on the cell surface and also to identify the similarities and differences between the lifestages. Across all samples and life stages 295 unique proteins were identified containing at

least 2 unique peptides ($p>0.05$). The number of proteins found in non-labelled trophozoite control was 7 and the number found in the non-labelled pseudocystcontrol was 5. For analysis, results for the multiple elutions for each trial were combined. Between the different biotinylated experimental trials there was a wide range in the number of unique proteins identified. In trial 1, 27 proteins were identified in the trophozoite sample and 29 in the pseudocyst sample. In trial 2, 276 proteins were identified in the trophozoite and 285 in the pseudocyst and in trial 3, 113 proteins were identified in the trophozoite sample and 135 in the pseudocyst. This does make comparing the expression between lifestages and the controls more difficult as the samples are so variable.

25 proteins were found in all three of the biotinylated trophozoite samples and all 25 were also found in at least two of the three biotinylated pseudocyst trials. 24 proteins were found in all three pseudocyst trials and, again, all were found in the trophozoite trials.

When comparing how many proteins appeared to be lifestage specific, in total, there were 12 proteins that were found in at least 1 trophozoite trial but in none of the pseudocyst trials (Table 3.5) and there were 21 proteins found in at least 1 pseudocyst trial but in none of the trophozoite trials (Table 3.6). Of those found in the trophozoite sample, V-ATPases (TTF81703) are known to be found in the plasma membrane [393] and the C2 [394] domain (TTF10759) is known to show selectivity for the cell membrane. However, many are proteins that confer generic functions within the cell or metabolism, such as 14-3-3 protein (TTF12165) [395] and NAD-dependent malic enzyme (TTF55649).

Protein ID	Product
TTF19128	hypothetical protein TRFO_32317
TTF12165	14-3-3 protein
TTF55649	NAD-dependent malic enzyme
TTF48506	hypothetical protein TRFO_11810
TTF57088	RNA-binding protein
TTF10759	putative C2 domain containing protein
TTF14285	UTP–glucose-1-phosphate uridylyltransferase family protein
TTF81703	V-type ATPase 116kDa subunit family protein
TTF16698	Carbamate kinase
TTF41230	RHO1-GTP-binding protein-rho subfamily of ras-like proteins
TTF27729	RNA-binding protein
TTF33240	glycyl-tRNA synthetase

Table 3.5: Biotinylated proteins found in Trophozoites and not Pseudocysts using label-free mass spectrometry. All proteins were found in at least 1 of the 3 trophozoite biotin trials but not in any of the pseudocyst trials nor either of the control samples. Samples were run on an Orbitrap mass spectrometer and results were analysed in Peaks [373].

Within the pseudocyst specific proteins, there are six hypothetical proteins and several enzymes,

such as tRNA ligase (TTF76542 and TTF59061) and oleate hydratase (TTF54486). Very few of the proteins seem to be specific to the *T. foetus* cell surface. However, there is an adhesin precursor (TTF65538) and a variable surface antigen like protein (TTF42565) which are highly likely to be cell surface expressed.

Protein ID	Product
TTF65538	adhesin AP65-1 precursor
TTF42565	Immuno-dominant variable surface antigen-like
TTF29274	hypothetical protein TRFO_21306
TTF61869	hypothetical protein TRFO_08502
TTF84159	long chain base biosynthesis protein 2a
TTF01232	DUF1846 domain-containing protein
TTF17659	Glucokinase 1
TTF44117	rubrerythrin family protein
TTF76542	serine–tRNA ligase
TTF81851	Dynamin central region family protein
TTF43479	beta subunit of citrate lyase
TTF00183	chaperonin containing TCP1 beta subunit
TTF16218	T-complex protein 1 subunit beta
TTF54486	oleate hydratase
TTF12141	hypothetical protein TRFO_21694
TTF12142	hypothetical protein TRFO_21694
TTF78733	hypothetical protein TRFO_27310
TTF70512	aldehyde dehydrogenase family protein
TTF32993	ornithine transcarbamylase
TTF59061	valine-tRNA ligase
TTF41913	hypothetical protein TRFO_42469

Table 3.6: Biotinylated Proteins Found in pseudocysts and not trophozoites using TMT mass spectrometry. All proteins were found in at least 1 of the 3 pseudocyst biotin trials but not in any of the trophozoite trials nor either of the control samples. Samples were run on an Orbitrap mass spectrometer and results were analysed in Peaks [373].

When all trials for each lifestage were combined, 70 proteins were found to be more abundant in the pseudocyst (fold change >2) (Table B3) 43 proteins were found to be more abundant in the trophozoite samples (Table B2). In the trophozoite samples, the majority (23), of proteins are ribosomal. This does not appear to be similar to the results of the label-free (Table 3.3) or TMT-labelled results (Table B1) there are also 4 hypothetical proteins (TTF56658), TTF08567, TTF27101 and TTF01106). There are also few clear cell-surface proteins apart from the C2 domain containing protein (TTF41309) and immuno-dominant variable surface antigen-like (TTF4080). This provides evidence that the labelling has not been specific to the cell surface. In the pseudocyst induced samples there is a much wider range of proteins than in the trophozoite preferentially upregulated samples. There is a thioredoxin protein (TTF22695), V-type subunit (TTF28277) and C2 domain containing protein (TTF10766) which suggest the cell surface has been labelled. However, the presence of several enzymes, such as catalase (TTF23521) and glutamate decarboxylase (TTF06177)

suggest other areas of the cell, such as the cytoplasm, have also been labelled.

When the abundance, in terms of peak area of both lifestages was compared, the top 50 abundant proteins were identified (Table B4). The most highly abundant protein was a thioredoxin precursor (TTF81436) followed by an adhesin precursor (TTF65227). There are tubulin, enolases and actin related proteins showing high levels of abundance. Tubulin, thioredoxin and enolases have all been found to be associated with the cell surface or organisms in varying capacities [396]. There are, again, many proteins that would not be expected to be associated with the cell membrane: ribosomal proteins, indole-3-pyruvate decarboxylase and ligases, showing that the cell labelling with biotin has not been specific to the cell surface.

3.3.6 Comparison of Cell Surface Predictions and Proteomics

Several of the proteins found in the biotin pulldown analysis were also found in the predicted cell surface proteome. Thioredoxin peroxidases were presentin the *in silico* network in Chapter 1, particularly in the largest cluster, and as stated, may have roles in protective responses [325]. Thioredoxins are also known to be found on the cell surface of trichomonads [325] providing evidence that the correct area of the cells have been labelled. Tubulin is involved in the formation of the cell membrane and so would be expected to be localised to the cell-surface, however, they would be expected to face inwards and would not be expected to be labelled. Profilins have be linked to membrane traffiking and signalling pathways between the cell membrane and cell cytoskeleton [397]. However, overall, there see to be a large number of general proteins found in the cell-surface biotinylated proteome that match general genes found in the transcriptome, for example, acetyltransferases and ABC transporters suggesting that is a more general proteome rather than a cell-surface enriched one.

3.3.7 Immunogenicity assays

The immunogenicity analysis consisted of three assays. Six microarray chips were each incubated with one of three sera, with two replicates of each serum, either pre-experimental infection (negative controls), post-experimental infection or post-natural infection (positives).

The images obtained the PEPslide Analyzer demonstrated a low background signal values with very weak interaction between the conjugate and the epitope, suggesting there is no interaction between the conjugate and the peptides.

The two replicates of each assay were combined to produce an average fluorescent response for each peptide spot. Distinct patterns can be seen between the positive and negative samples. The pre-immune sera showed high fluorescence at random positions and no fluorescence at others (Figure 3.12). The post-immune experimentally infected samples produced higher fluorescence than the negative control and naturally infected samples (Figure 3.13). The naturally infected samples had lower fluorescence than experimentally infected animals and in some cases the raw values for peptides were similar to those of the negative controls (Figure 3.14).

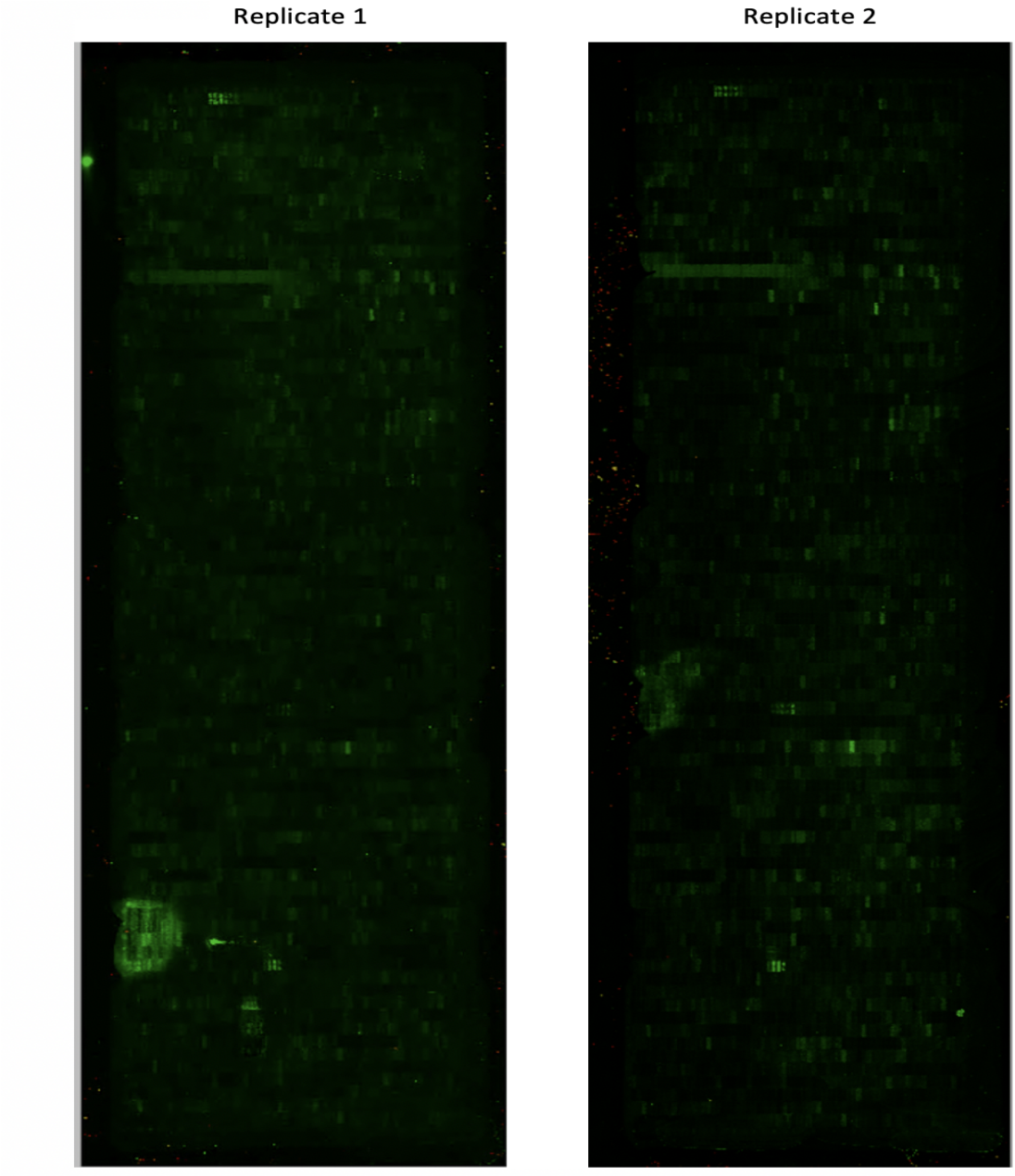

Figure 3.12: Spot intensities from *T. foetus* peptide microarray incubated with pre-infection sera from experimentally infected cattle (negative controls). The images were obtained from the PEPslide Analyzer. Each spot corresponds with a peptide. Two replicates were performed.

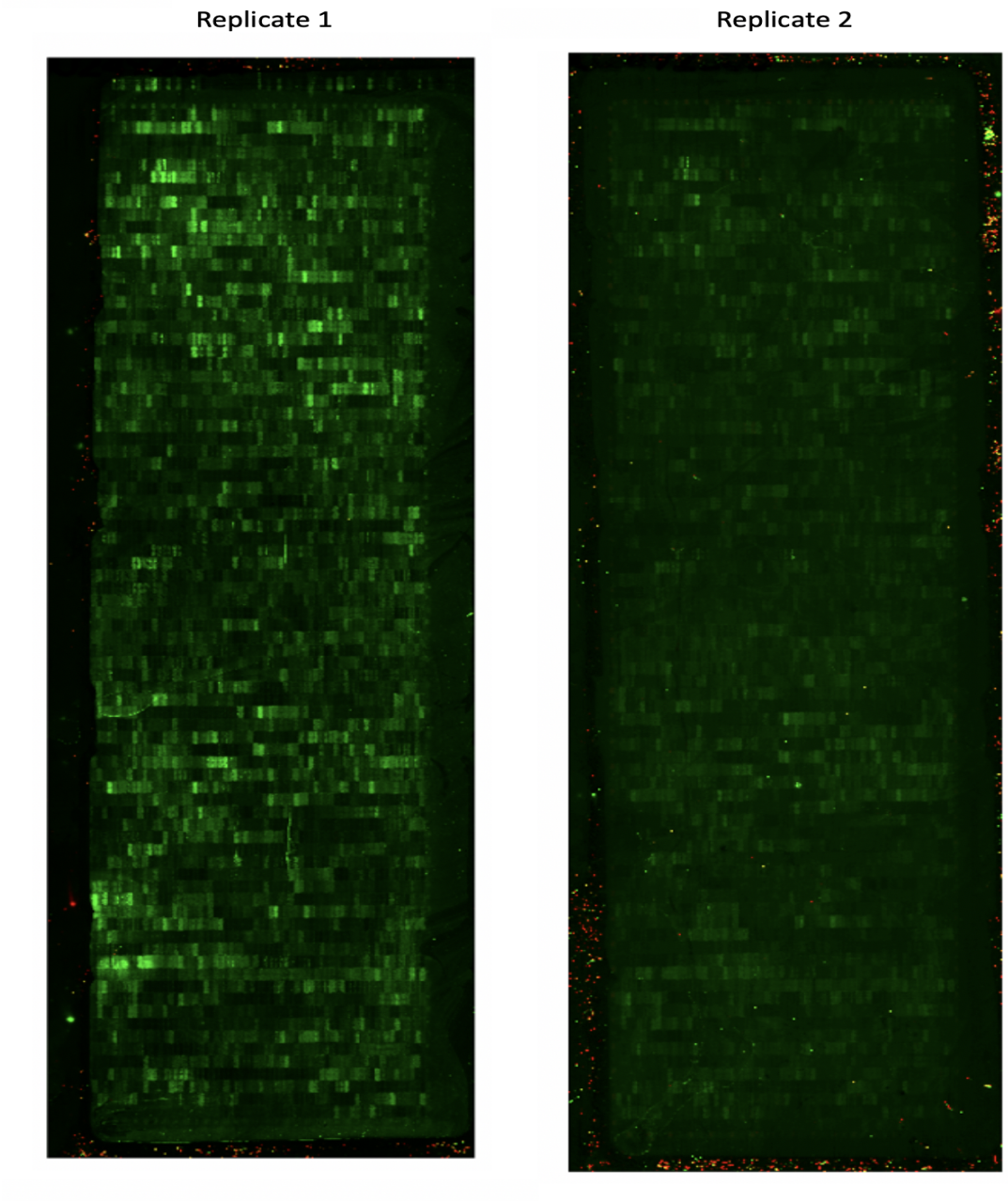

Figure 3.13: Spot intensities from *T. foetus* peptide microarray incubated with post-infection sera from experimentally infected cattle (positive samples) The images were obtained from the PEPslide Analyzer. Each spot corresponds with a peptide. Two replicates were performed.

Figure 3.14: Spot intensities from *T. foetus* peptide microarray incubated with post-immune sera from naturally infected cattle (positive samples) The images were obtained from the PEPslide Analyzer. Each spot corresponds with a peptide. Two replicates were performed.

The raw intensity values were analysed using limma in R. When the pre-immune (negative control) and post-immune experimental sera signal intensities were compared, the analysis identified 1532 epitopes (padj <0.05 and LogFC>2) that has significantly high fluorescence in the post-immune experimentally infected sample. Likewise, when the intensities for the negative control and naturally

infected animal samples were compared 791 epitopes were identified (padj <0.05 and LogFC>2). There were fewer peptides identified in the naturally infected sera than the experimentally infected sera, 1532 and 791 peptides were identified in experimentally and naturally infected sera respectively compared to the pre-infection sample.

Proteins were identified that induced a strong immune response from the post-infection sera. The number of significant immunogenic epitopes for all 51 proteins in the microarray, and the top twenty proteins by epitope number are listed for both natural and experimental infections, in Table 3.7. Among proteins that are highly responsive to both natural and experimental post-immune serum, TF11197 was the protein with most immunogenic epitopes (167), followed by TTF53402 (163) and TTF00910 (151).

Protein ID Experimental Infection	No. of Epitopes	Protein ID Natural Infection	No. of Epitopes
TTF00910	99	TTF00910	52
TTF03161	44	TTF03161	14
TTF09063	48	TTF09063	25
-	-	TTF09721	20
TTF11197	88	TTF11197	79
TTF12012	75	-	-
TTF13670	38	TTF13670	24
TTF13877	52	TTF13877	25
TTF14901	84	TTF14901	62
TTF16447	31	TTF16447	16
TTF19157	35	TTF19157	30
TTF34783	27	TTF34783	26
TTF38179	85	-	-
TTF44625	36	-	-
TTF44949	47	TTF44949	37
-	-	TTF49921	28
TTF53402	105	TTF53402	58
TTF55982	35	-	-
TTF56587	43	TTF56587	13
-	-	TTF59933	20
-	-	TTF67688	26
TTF72966	63	TTF72966	62
TTF73824	59	TTF73824	32
TTF75485	46	-	-
-	-	TTF75903	14

Table 3.7: The protein IDs and number of significant epitopes of the proteins with the largest 20 numbers of expressed epitopes for the experimentally infected cows and naturally infected cows

Of the top 20 proteins by epitope number, 15 were identified in both the experimental and natural infection assays (Table 3.7. In some cases the number of epitopes was similar for sera from both naturally infected and experimentally infected cattle, such as TTF34783 and TTF72966, however

in many others, such as TTF53402 and there were very large differences with 105 compared to 58. Overall, there was a positive correlation between the number of significant epitopes in proteins recognised by serum from naturally infected cows compared to experimentally infected cows (Figure 3.15) with an R^2 of 0.5784. Therefore, those proteins that had a large number of significant immunogenic epitopes recognised by in the experimentally infected samples also tended to have larger numbers in the naturally infected samples, although there are several outliers. These were significantly higher than the pre-infection sera.

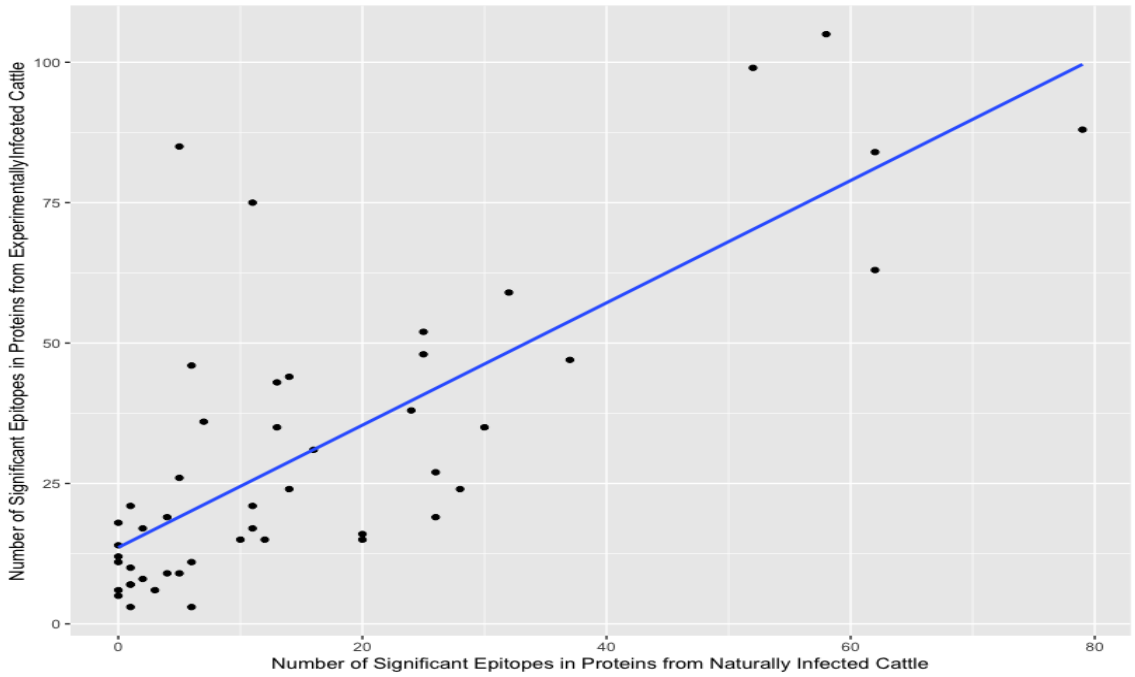

Figure 3.15: Correlation of the number of significant immunogenic epitopes in peptide array proteins using serum from experimentally infected and naturally infected cattle. Significance was calculated using padj <0.05 and LogFC>2. Graph was produced using ggplot2 in R and a linear regression line was added with an adjusted R^2 value of 0.5784.

While the same proteins appear to contain the most immunogenic epitopes detected by both types of infected serum, this does not guarantee that these proteins elicit a dominant response, since no account has been taken of the magnitude of the response. The fold change in fluorescence intensity of each peptide, relative to negative controls, was used to estimate the abundance magnitude of the antibody. When the fold changes in fluorescence for each peptide probes with sera from experimentally and naturally infected cattle were compared, there was no obvious correlation (Figure 3.16). This shows that a strongly immunogenic epitope, detected by antibodies from experimentally infected animals does not necessarily mean the same epitope will have a high level of antibody in

sera from naturally infected animals and vice versa.

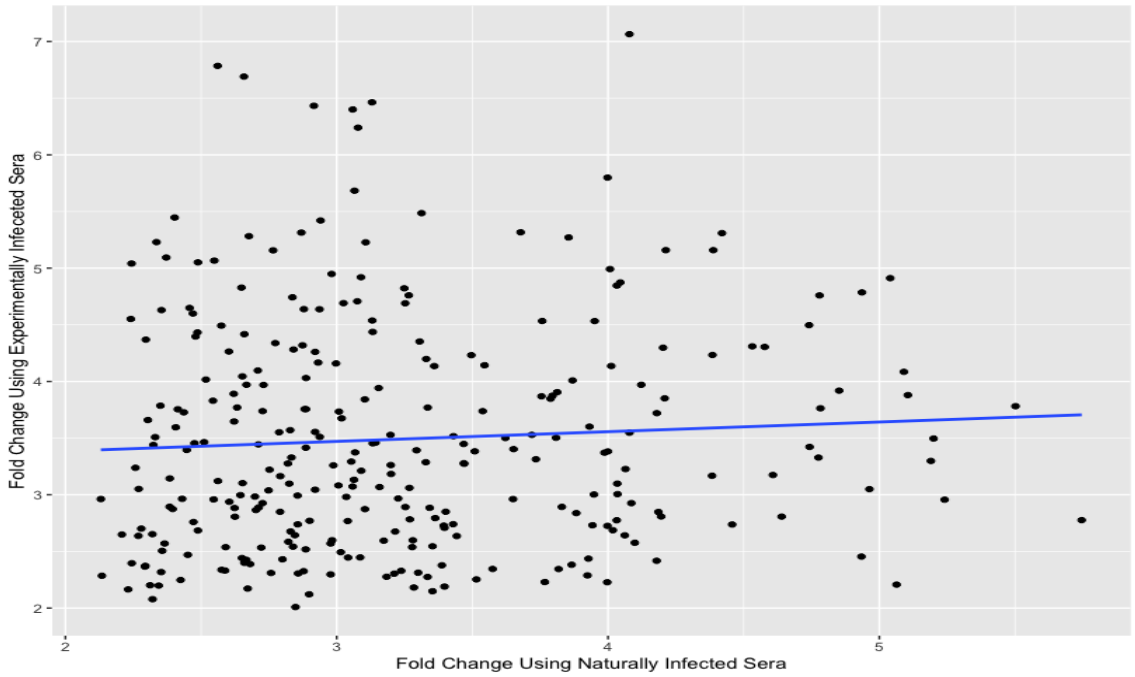

Figure 3.16: Comparison of the fold changes of significantly immunogenic epitopes expressed in the peptide array. Fold change was calculated using limma in R, comparing the spot intensities of the post-infection sera samples to the negative controls. The fold changes of epitopes found to be expressed in both post-infection trials were then compared. Graph was produced using ggplot2 in R and a linear regression line was added with an adjusted R^2 value of 0.0005263

When the fold change of the epitopes recognised by antibodies were compared and the adjusted P value (Padj) calculated ,the results were markedly different between the epitopes expressed from the experimentally infected samples (Figure 3.17) and those from the naturally infected samples (Figure 3.18). The epitopes with the highest fold change were ranked, as were those with the lowest Padj value. The 200 epitopes with the highest combined score, i.e. high fold change and low Padj were used for the correlation as these showed the most significant and high intensity expression. The epitopes from the experimentally infected sera showed a much weaker correlation than those from the naturally infected sera samples, with adjusted R^2 values of 0.01022 and 0.5421 respectively. There is a number of epitopes from the experimentally infected samples that have a log fold change over 4.5 but a Padj over 0.002 meaning they are more likely to have occurred by chance, however, the vast majority of the epitopes have a Padj under 0.02. For the naturally infected samples, there is a stronger correlation between the log fold change and Padj, however, the spread of Padj values is much wider, with the majority falling between 0.002 and 0.007. This shows that whilst the correlation between values is lower for the epitopes from experimentally infected samples, the

significance is higher. Furthermore, the fold change values for the epitopes from experimentally infected sera are higher than the naturally infected sera, with the vast majority ranging from 4 to 7.5 compared to under 5.

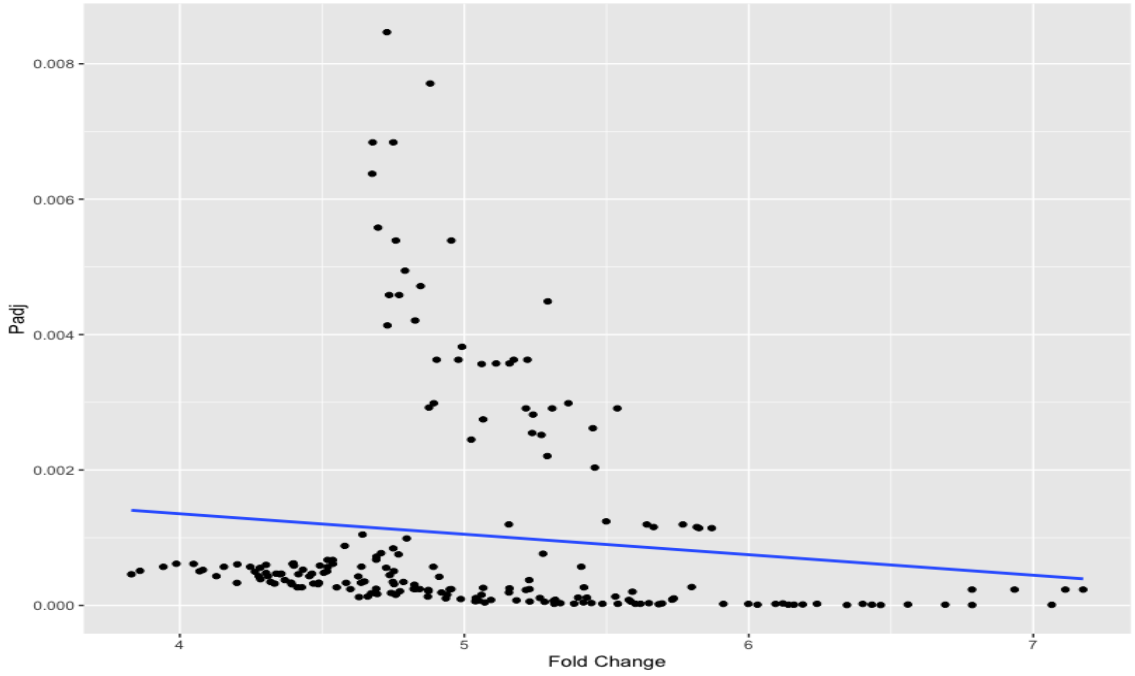

Figure 3.17: Correlation of the log fold change and adjusted P value of epitopes from the peptide array expressed using sera from experimentally infected cattle. The epitopes with the highest fold change were ranked, as were those with the lowest Padj value. The 200 epitopes with the highest combined score, i.e. high fold change and low Padj were used for the correlation. Graph was produced using ggplot2 in R and a linear regression line was added with an adjusted R^2 value of 0.01022

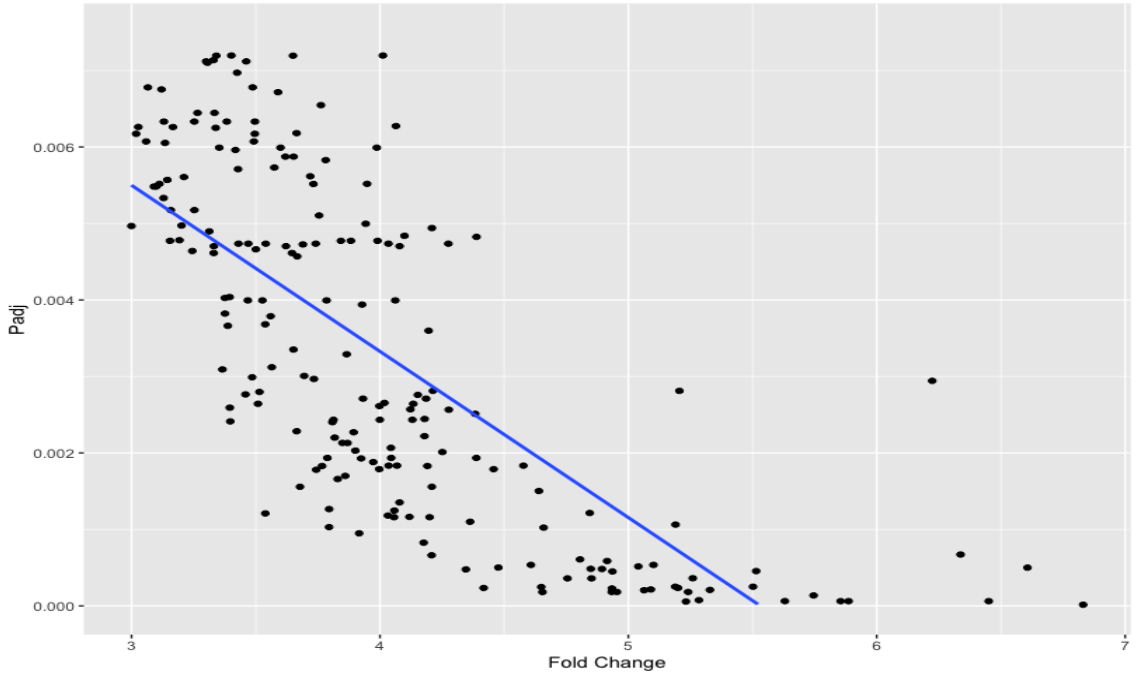

Figure 3.18: Correlation of the log fold change and adjusted P value of epitopes from the peptide array expressed using sera from naturally infected cattle. The epitopes with the highest fold change were ranked, as were those with the lowest Padj value. The 200 epitopes with the highest combined score, i.e. high fold change and low Padj were used for the correlation. Graph was produced using ggplot2 in R and a linear regression line was added with an adjusted R^2 value of 0.5421

The maximum intensity of each protein was calculated using the significant epitope from each protein which had the highest fold change relative to the negative control. This was then corelated against the number of significantly responsive epitopes found in the protein overall. This was performed for both the experimental sera and the natural sera (Figures 3.19 and 3.20). Overall, for both sera types there was a positive correlation between the log fold change of the most reported epitope in a protein and the number of significant epitopes in the protein. There were 9 proteins for the experimental sera samples that had over 50 significant epitopes and a maximum intensity of over 4.5 times that of the negative control (Figure 3.19 and Table 3.8), whereas, there were 5 proteins for the natural sera that fulfilled the same criteria (Figure 3.20 and Table 3.9). All 5 proteins identified in the naturally infected sample were also identified in the experimentally infected sera: TTF00910, TTF11197, TTF14901, TTF53402 and TTF72966.

Protein ID	No. of Significant Epitopes	Maximum Intensity
TTF38179	85	7.17
TTF12012	75	7.07
TTF73824	59	6.68
TTF11197	88	6.55
TTF72966	63	6.34
TTF00910	99	6.09
TTF13877	52	5.91
TTF14901	84	5.33
TTF53402	105	5.06

Table 3.8: The protein IDs, number of significant epitopes and maximum intensity of those epitopes from Figure 3.19. Proteins were included if their maximum intensity/log fold change was higher than 4.5 and the protein possessed over 50 significant epitopes. Maximum intensity was calculated using the epitope with the highest log fold change in the protein relative to the negative control when sera from experimentally infected animals was used.

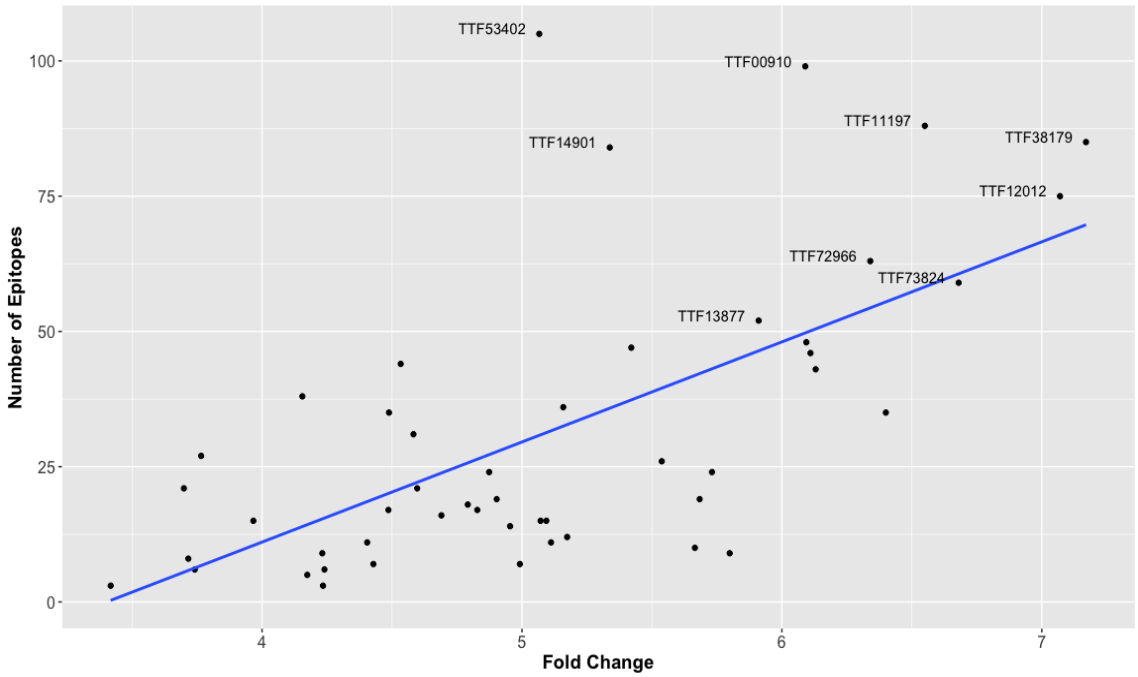

Figure 3.19: The Correlation of the number of significant epitopes per protein in the microarray and the maximum intensity of the protein. Maximum intensity was calculated using the highest log fold change of an epitope in the protein relative to the negative controls. For this experiment, sera from experimentally infected cattle were used. The IDs of those proteins that had a maximum intensity/log fold change over 4.5 and contained over 50 significant epitopes were added. A regression line has also been added. There was a positive correlation between the values, with an R^2 value of 0.4068 and Pearson correlation coefficient of 0.6378 (95% confidence).

Protein ID	No. of Significant Epitopes	Maximum Intensity
TTF00910	52	5.99
TTF11197	79	5.74
TTF14901	62	5.23
TTF53402	58	5.06
TTF72966	62	4.84

Table 3.9: The protein IDs, number of significant epitopes and maximum intensity of those epitopes from Figure 3.20. Proteins were included if their maximum intensity/log fold change was higher than 4.5 and the protein possessed over 50 significant epitopes. Maximum intensity was calculated using the epitope with the highest log fold change in the protein relative to the negative control when sera from naturally infected animals was used.

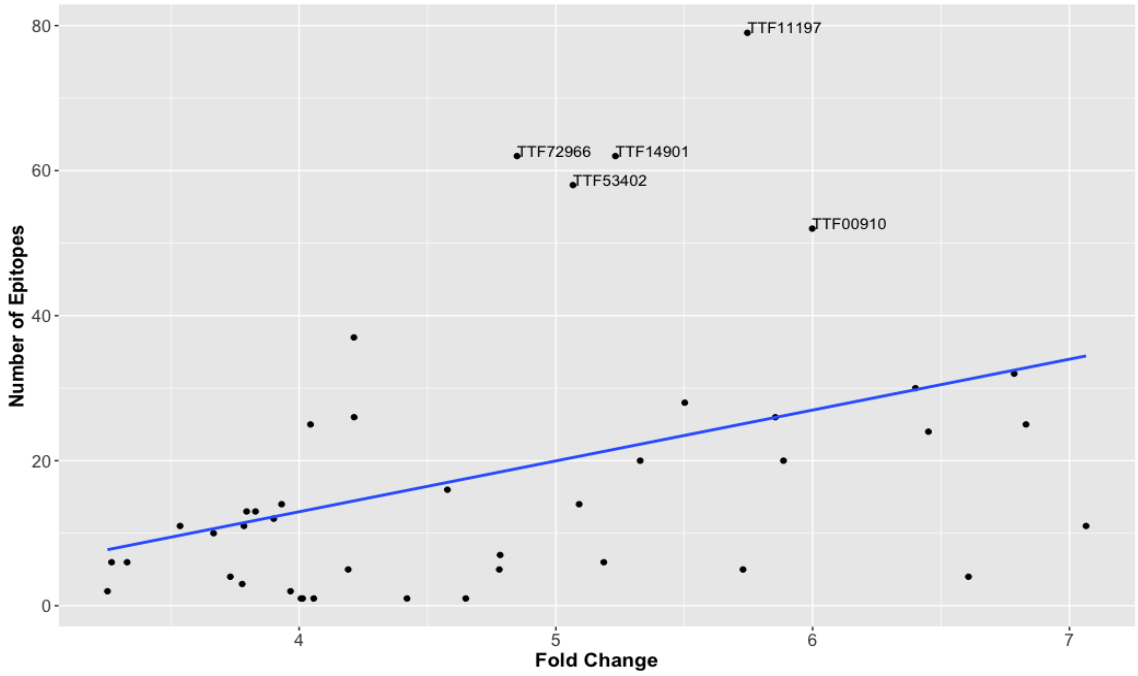

Figure 3.20: The Correlation of the number of significant epitopes per protein in the microarray and the maximum intensity of the protein. Maximum intensity was calculated using the highest log fold change of an epitope in the protein relative to the negative controls. For this experiment, sera from naturally infected cattle were used. The IDs of those proteins that had a maximum intensity/log fold change over 4.5 and contained over 50 significant epitopes were added. A regression line has also been added. There was a weak positive correlation between the values, with an R^2 value of 0.1576 and Pearson correlation coefficient of 0.397 (95% confidence).

3.3.8 Comparison of Peptide Array Results and *in silico* Predictions

When compared to the network in Chapter 1, TTF11197 and TTF00910 were both found in the same cluster, along with TTF72966 which was also a peptide strongly recognised by antibodies in both the experimentally infected and naturally infected samples (Figure 3.21). All three proteins

were found to have a high number of significant epitopes: 88, 99 and 63 for TTF11197, TTF00910 and TTF72966 respectively in the experimental sera sample and 79, 52 and 62 in the natural sera sample, and a high fold change relative to the negative control, over 4.5 times in all cases (Tables 3.8 and 3.9). The cluster itself contains 25 genes, all of which are *T. foetus* specific and contained four other proteins that contained signal peptides. This family of genes itself contains *T. foetus* specific proteins only, at least three of which are immunogenic.

TTF53402, which had 263 immunogenic peptides was part of a triplet family which contained another *T. foetus* specific gene and a *T. vaginalis* specific protein and TTF44949 was a member of a doublet, the other member of which was a *T. vaginalis* gene. The only protein that was highly immunogenic, with over 30 peptides in both naturally and experimentally infected samples and found in the largest cluster was TTF73824. Of the proteins found only in the post-immune experimentally challenged serum, two of them are found in the large cluster: TTF44685 and TTF45090. TTF23499 is found as a double and TTF23900 is found in the Clan-SC cluster.
There are 11 of the top 20 expressed epitope peptides found in the network when the experimentally (Figure 3.22) infected sera was used compared to 13 when the naturally infected sera was used suggesting there were more single copy genes that had immunogenic epitopes to the experimentally infected cattle sera.

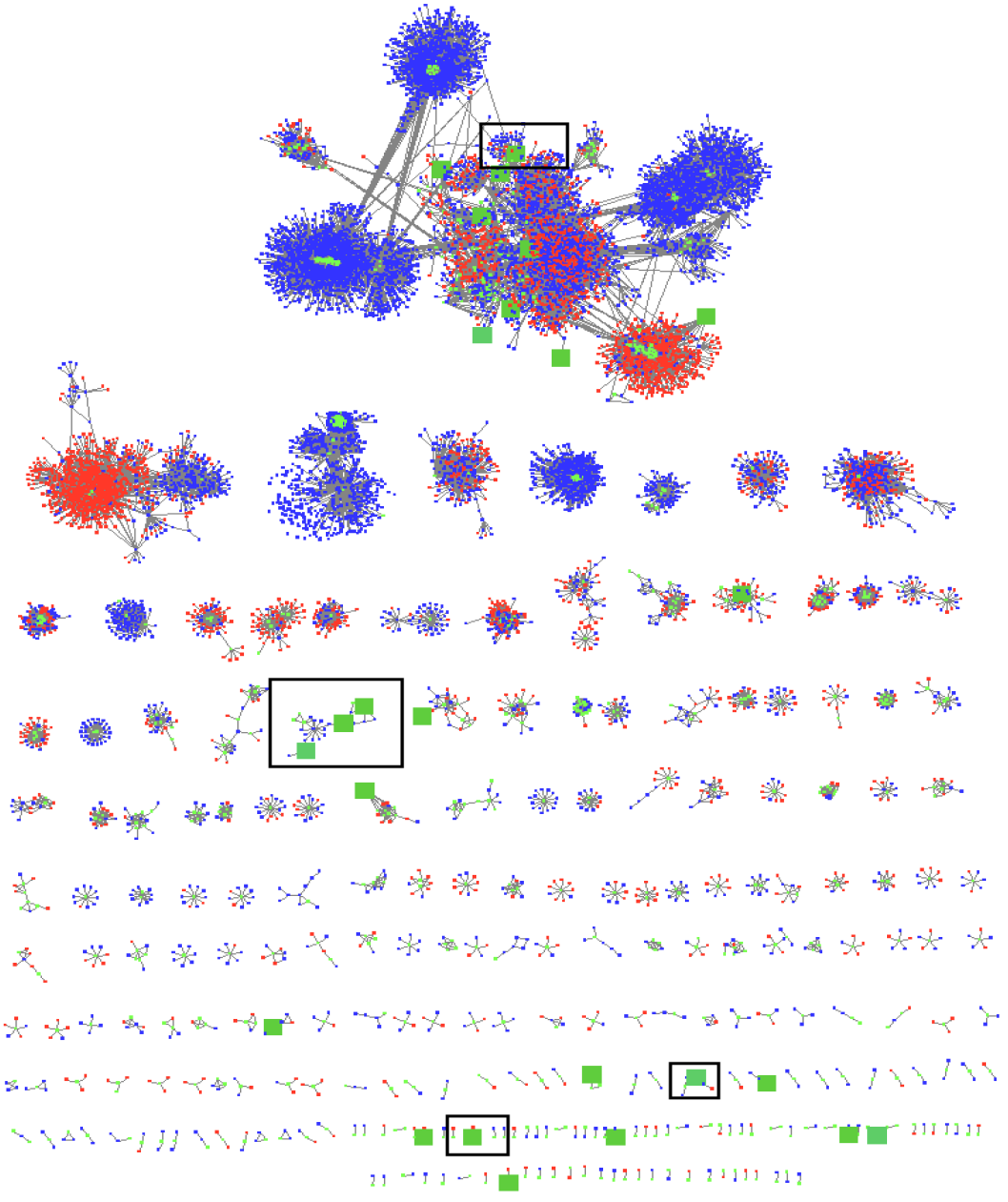

Figure 3.21: Network showing *T. vaginalis* (red) and *T. foetus* (blue) genes produced in Cytoscape. All members that contain a signal peptide are highlighted in green. Additionally, peptides found in the peptide array are highlighted as large green squares. Highly immunogenic peptides identified in the peptide arrays when both sera types were used: experimentally infected and naturally infected animals are highlighted using a box. These are TTF11197, TTF00910 and TTF72966 found all found in the same cluster; TTF53402 found as a triplet, TTF73824 found in the largest cluster and TTF44949 found in a doublet. All of these proteins had high numbers of significant immunogenic epitopes and were found in both infected sera samples.

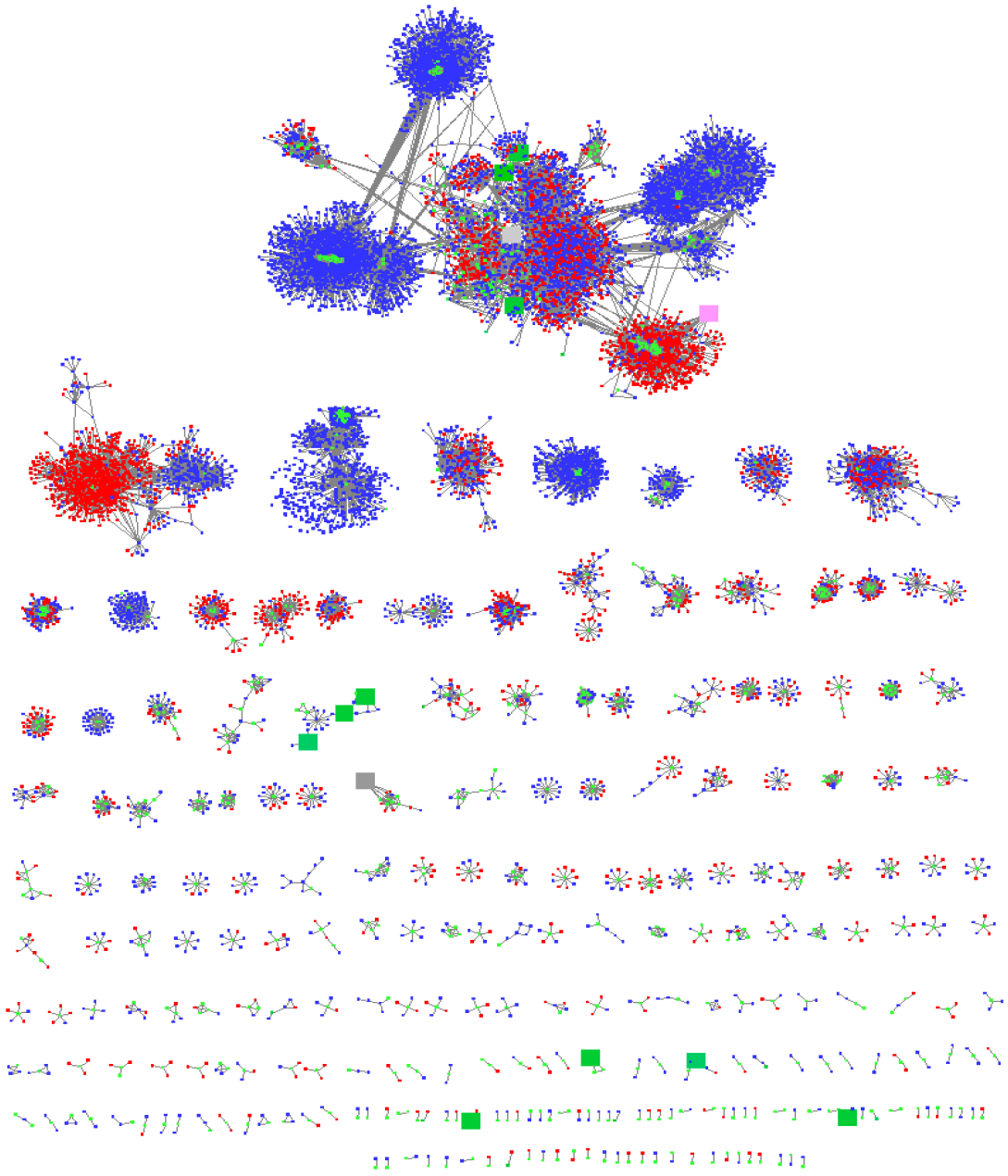

Figure 3.22: Network showing *T. vaginalis* (red) and *T. foetus* (blue) genes produced in Cytoscape. All members that contain a signal peptide are highlighted in green. Additionally all of the proteins with the 20 highest epitope numbers of expression that are also present in the network, and show significant immunogenicity when sera from naturally or experimentally infected animals is used are highlighted. Those that are found in the top 20 of both the experimentally and naturally infected samples are highlighted as green squares, those found only in the experimental sera samples were highlighted in pink and those in natural sera grey.

3.4 Discussion

In this chapter, the cell-surface of both *T. foetus* life-stages were labelled using biotin and extracted by way of a biotin-streptavidin pulldown. Proteomic analysis was performed by TMT mass spectrometry (Figure 3.23). Additionally, proteins from the *in silico* cell surface proteome described in Chapter 1 were analysed for their immunogenicity using a peptide microarray, probed with sera from experimentally and naturally infected cows. This has identified numerous parasite antigens that elicit consistently strong antibody titres, a feature that indirectly corroborates their predicted cell-surface localisation, and is a key feature of a good vaccine candidate.

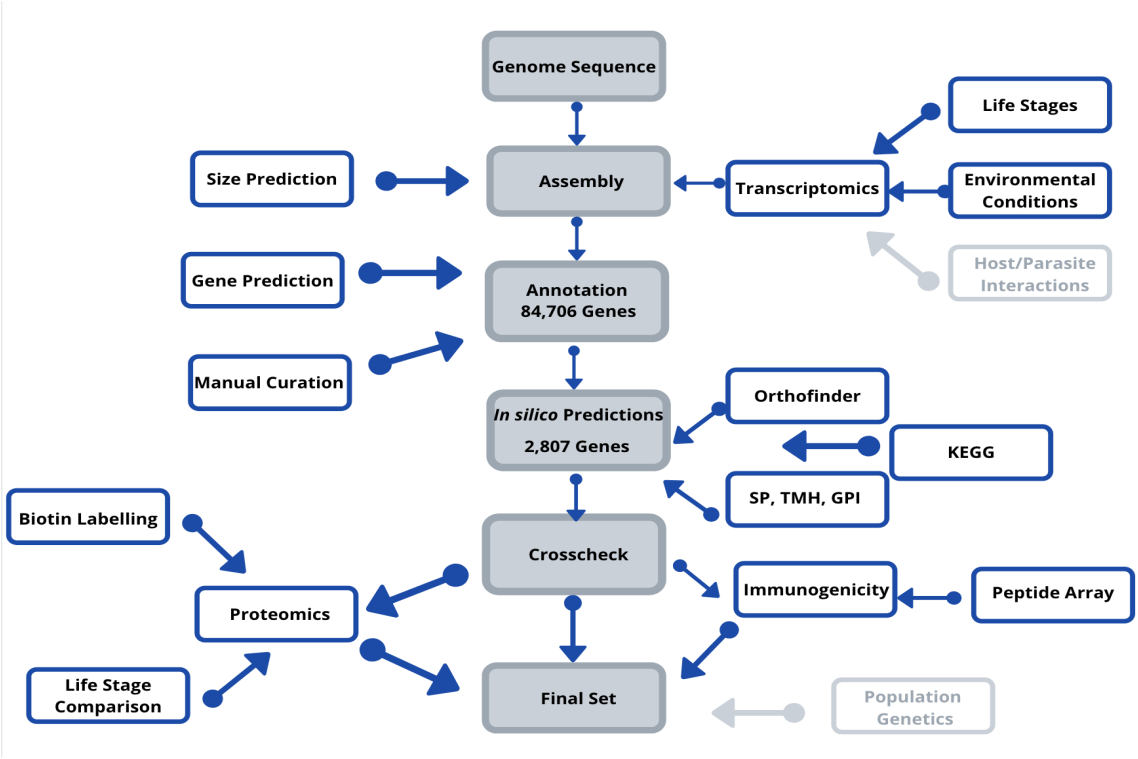

Figure 3.23: Reverse Vaccinology Project Flow Diagram. The sections highlighted are those that have been completed over the course of this chapter and those that are pale are to be completed in subsequent chapters.

3.4.1 Cell Surface Proteins

The aim of this chapter was to find immunogenic peptides that could be used as vaccine candidates, however this pipeline excludes glycolipids, which could be vaccine candidates [398] [399]. The peptides themselves may also be subjected to PTMs which could affect their structure or function. The pipeline used to select the peptides themselves also removes any peptides that do not have a TMH, meaning any that are purely secreted will not feature on the array. These secreted proteins may have an key role in virulence, as is the case for some *T. vaginalis* secreted proteins such as β-amylases [400].

In the case of a rabies vaccine [401], attempts were made to create a DNA vaccine using a rabies glycoprotein and four constructs were made using the rabis-virus glycoprotein ectodomain. One containing the signal peptide sequence and transmembrane domain, one containing the SP but no TMH, one containing a TMH but no SP and one containing neither the SP or TMH. When a prime-boost strategy was performed, the mice primed with the construct with both the SP and TMH produced the highest antibody titres when challenged, followed by mice primed using the construct with a TMH but no SP [401]. These results show that priming with the rabies construct containing the full glycoprotein and TMH provides the best protective response, showing the importance of the TMH being present in the immunogenic peptide. Furthermore, both SP and TMH have been found to have high epitope densities [402] relative to sequence length and can show strong MHC binding, further showing their importance in epitope selection and vaccine design.

There are some issues with using signal peptides as a key factor in choosing the peptides. The translated proteins may not have the correct start codon or may use non-standard SPs. If the original annotation is not correct then the TMH could be missidentified as a GPI anchor or signal peptide, however, choosing proteins that contain both SP and TMH with clear sequences would help to minimise this.

3.4.2 Proteomic Differences Between Trophozites and Pseudocysts

No strong key differences were seen between the trophozoite and pseudocyst samples, both biotinylated and non-biotinylated. No clear distinctions were seen in the IFAs, banding patterns of the Western blots or in the proteomic analysis. This is potentially an issue of inducing the pseudocysts themselves as the most successful way we have found to induce them is by subjecting the *T. foetus* cells to low temperatures (4°C). Keeping all reagents and preparation stages at this temperature

can prove difficult and the optimal temperature for biotinylation of the cells is 4°C. This means that both cell samples are subjected to temperature changes which can cause morphology changes and therefore mixed populations rather than pure trophozoites or pseudocysts, thereby skewing the results.

Very few proteins were seen that would have been expected to be in high abundance in any cell, such as tubulin or cysteine proteases. There was also very little similarity with the stage specific transcriptome results, such as the heat shock proteins. There were many ribosomal proteins, along with many enzymes, which suggests it was not only the cell surface that was labelled. There were thioredoxin proteins and elongation factors found in relatively large numbers in the biotinylated trials which were not seen in the top 10 up or down regulated transcriptomes for any of the different environmental conditions. It may be that the proteomics methods were not sensitive enough to pick up many of the proteins, but it would still be expected than some, such as the myb-like binding domains or N-acetyl-transferases would be seen in some number, particularly if more than just the cell surface was labelled.

Between the TMT labelled and the label-free proteomics experiments there were few similarities between the responses. This may be in part tat the label-free experiments were only using one replicate per life stage rather than three in the labelled experiment and the fact that there were a significantly smaller number of proteins seen, for example only 7 significant proteins found in the label-free sample. This makes comparisons between the two more difficult. For the low numbers of proteins that are seen, the only proteins that are found in both experiments are hypothetical proteins and ribosomal proteins.

In regards to the biotin proteomic results, there was only one sample with a significant number of peptides and so there was no replication due to the other two samples not being high enough quality or having a large number of peptides. Due to this the samples were pooled and then the lifestages compared. It was possible to obtain abundance values for the two lifestages generally, however, a large number of proteins from each sample were only found in one of the three trials, reducing the overall reliability.

3.4.3 Peptide Array

Overall, 1532 and 791 peptides were identified using experimentally and naturally infected sera respectively compared to the pre-infected (negative control) sample. Of the top 20 proteins that had the highest number of significantly immunogenic epitopes, 15 were shared in both infected

trials. There were lower spot intensities using sera from naturally infected animals overall when compared to the experimentally infected animals. This may be due to lack of stimulation in the naturally infected animals, or the challenge itself may have been lower, ie. fewer parasites and so the immune response may have been lower. The response itself may have been short -term rather than long lasting. Furthermore, The strength of binding between the antibody and some peptides may have been low [403] even if they are immunogenic. This may have been due to the duration of the infection or there may be conformational epitopes that are not present in the microarray.

This may be due to the timing post infection of the samples being taken as different antibodies may be expressed at different stages or, as the naturally infected animals were infected prior to the experimentally infected animals the immune response may be lower and so key epitopes may not be expressed in the same way. Other explanations might be that the experiments were done on calves, whereas the natural infections came from adults, leading to difference in immune response. Also, the natural infections may have had lower parasitaemia than experimental infections, causing the lower intensity responses.

Other veterinary vaccines have also been using peptides as antigens [404], for example against foot and mouth disease (FMDV) and canine parvovirus [404]. Peptide vaccines have also been used in attempts to combat Porcine circovirus type 2 (PCV2) [405] and it was demonstrated that, when liposomes were used as adjuvants, the peptides could stimulate both humoral and cellular immune responses [405].

Of the proteins identified in the peptide array, TTF11197, TTF00910 and TTF72966 appear to be interesting as they show significantly high responses when probed with both natural and experimental sera and they are found within one *T. foetus* specific gene family (Figure 3.21). They all contain a signal peptide and transmembrane domain and other members of the same gene family also contain these signal peptides. TTF73824 is found in the largest cluster, that contains the BspA family and cysteine proteases, genes that are known to be involved in host parasite interactions. All proteins generate strong B-cell immunological responses, however, a B-cell response alone does not always necessarily confer full protective immunity [406] and so testing for T memory cell responses would also be crucial.

3.4.4 Future work

In terms of future work using these same approaches, further peptide arrays could be performed. These could involve proteins from several different strains and could identify if there are differences

in immunogenic proteins across the species. Arrays could also be performed using sera from animals of a different sex, i.e. bulls, as *T. foetus* affects the genders differently and so different peptides may be immunogenic. The arrays could be performed using sera from different time points to provide an insight into the progression of the disease over time and if different antigens are expressed and how long antibody responses last for. These arrays could also be used to identify targets for diagnostic assays as there seem to be reliable markers for infection and could potentially show not only if the animal is infected, but also how far along the disease is in its progression. The experiment could also be repeated on animals of different ages to identify cow lifestage specific immune responses.
The epitopes examined in these experiments were B-cell epitopes, and so looking also at T- cell epitopes to examine the cell-mediated responses would also be a next step. Furthermore, future work could involve labelling different strains of *T. foetus* as there may be differences in cell surface composition, e.g. adhesins, as some strains appear to be more cytotoxic than others and have higher levels of adherence. This could also be applied to strains that infect cats and may be a potential diagnostic strategy for feline infections.

3.4.5 Conclusion

This chapter has examined the immunogenicity of 50 putative cell surface antigens discovered in Chapter 1 and identified a cohort eliciting consistently strong immune responses *in vivo*. Some of these antigens belong to parasite-specific gene families that may represent important features of the *T. foetus* cell surface architecture. The task now, before reaching a final selection of the best vaccine candidates, is to examine the expression of these antigens in an *in vitro* disease model (Chapter 4) and the natural polymorphism in their protein structures (Chapter 5). Although these antigens may be strong immunogens, only when they are proven to be robustly expressed and structurally invariant can they be considered reliable candidates for vaccination.

Chapter 4

Transcriptomic Analysis of Parasite Gene Expression in a Host Co-culture Infection Model

This chapter describes a model of *T. foetus* infection in the form of parasite culture on a host cell monolayer. Growth of five strains of *T. foetus* (Belfast, KV-1, BP-4, UT and DK2) and one strain of *T. mobilensis* in the presence of an MDCK (canine kidney) monolayer is measured, and the rates of adhesion and cytotoxicity among the different Tritrichomonas strains is compared. Thereafter, it also describes a transcriptomic analysis of parasite gene expression (strain DK2) during MDCK co-culture. Gene expression profiles of parasites in various degrees of association with the monolayer are compared to axenic parasite cultures to identify those gene transcripts that are significantly more abundant during host cell interaction, and which therefore may be considered important for infection *in vivo*.

4.1 Introduction

The interactions between *T. foetus*, and indeed any parasite, and its host are crucial to its pathogenicity and the clinical signs it produces. If the proteins used in the *T. foetus* vaccine are expressed predominately in the presence of the host, targeting these has the potential to greatly reduce the virulence of the parasite. Successful vaccines target abundant antigens that are reliably expressed at

the host-pathogen interface. Discovering these for *T. foetus* requires an analysis of gene expression using a relevant model of the infection interface.

4.1.1 Cell Monolayers

Monolayers are tissues that are one-cell thick. They make up many components of multicellular organisms [407], such as the intestinal epithelia, and fulfil roles in development and protection [407]. Monolayers can also have different physiological characteristics to single cells, for example, Harris *et al.* (2012) [407] determined that monolayers have twice the elasticity of single cells. They also have the addition of cell-cell interaction components, such as tight junctions and adherent junctions [408] which would also be found *in vivo*. Therefore, when considering parasite gene expression, and so modelling all of these aspects is likely to give more accurate results than working with single cells. The benefit of monolayers is that they consists of living cells and so can mimic the responses of an organism. Furthermore, interactions between parasites and host cells often occurs on monolayers *in vivo*, such as in *T. vaginalis* infections of the urogenital tract.

Monolayers have long been used as a way to model host-parasite interactions *in vitro*. For example, human corneal monolayers have been used to model *Streptococcus aureus* keratitis in humans [409] and biofilms have been used to model periodontal disease [410]. In the case of intestinal epithelial monolayers, both luminal and basal sides can be examined and screened [408] if a membrane or scaffold is used. Monolayers have also been used to further the understanding of *Cryptosporidium* [411] and *T. vaginalis* [412]. They have allowed for host-parasite interactions to be modelled in mice. They have been used in numerous binding and cytotoxicity studies [413] [414], for example in the binding of feline *T. foetus* to IPEC-J2 (porcine) intestinal epithelial cells to determine their pathogenic effect [413].

4.1.2 Host-trichomonad Cellular Interactions

The mechanisms of host-trichomonad interactions for *T. foetus* or *T. vaginalis* are poorly understood. It has been postulated that *T. vaginalis* adheres to the host cells via adhesin proteins [67], however, the exact identity of these is still unknown. Proteomic analyses have identified secreted proteins, in addition to cell surface associated proteins that may play a role in the interactions [67]. When *T. vaginalis* adheres to the host cells, there is typically a developmental change from their trophozoite form to pseudocysts [67]. As stated in Chapter 1, there are three classes of molecules that have been postulated to be involved in host-cell adhesion in *T. vaginalis*: Lipophosphoglycan

(LPG), metabolic proteins and membrane proteins [67].

LPG is a surface glycan of the *T. vaginalis* glycocalyx [67]; surface carbohydrates such as galactose are involved in the parasite virulence as binding studies show that more virulent *T. vaginalis* phenotypes bind to agglutanin more than less virulent strains. This shows that differences in phenotype can greatly impact *T. vaginalis* virulence. Currently, the structure of LPG is not resolved but contains many poly-N-acetyllactosamine repeats that may be important [415] [416].

Metabolic proteins-adhesion proteins have been identified that are involved in carbohydrate metabolism and are often found in the hydrogenosome, such as AP65, AP33, AP51 and AP120 [162]. However, these same proteins have also been reportedly found on the cell surface [163] [162], meaning they may have duel functions as both metabolic proteins and adhesins. Furthermore, some of these proteins, AP51 and AP65, have also been shown to bind non-specifically, such as to haem groups [417] and, as yet, no specific binding partners have been found [67].

T. foetus is known to adhere to monolayers and it was thought that this adherence may also lead to cytotoxicity [38]. As stated in Chapter 1, laminin and fibronectin may mediate this host-cell adhesion. Both *T. suis* and *T. foetus* were found to adhere to mucus, mediated by lectin binding [111] [38]. Extracellular proteases, capable of cleaving fibronectin, are also thought to be involved in the host-parasite interactions and pathogenesis and may damage host cell integrity [38].

Membrane proteins-of the proteins identified in the *T. vaginalis* genome, serine and cysteine proteases were implicated in pathogenesis [67] due to their similarity to virulence-associated proteins from *Leishmania*, and they may have roles in degrading the host cells cytoskeleton or extracellular matrix. Immunodominant variable surface antigens were also identified which may be involved in the attachment of the *T. vaginalis* parasites to the host epithelia [67]. Furthermore, BspAs were identified, which have also been thought to be involved in virulence [67] [167] [327].

4.1.3 Host-cell Adhesion and Cytotoxicity in Trichomonads

It is known that infection by trichomonads can cause a range of clinical signs. Numerous studies have documented that trichomonads, particularly *T. foetus* and *T. vaginalis*, can adhere to host cells and can display cytotoxic activity [69] [78] [112] [38] [155] [33]. Tolbert *et al.* (2014) [413] examined the pathogenic effect of *T. foetus* parasites, acquired from cats, on intestinal epithelial cells and looked at the effect of adhesion on the pathogenicity of *T. foetus* cells. They used an *in vitro* monolayer co-culture system and demonstrated that direct contact of *T. foetus* and host

cells triggered apoptosis of the epithelial cells. These effects were also shown to be largely due to secretion of *T. foetus* cysteine peptidases, which also promoted cellular binding [413].

Tolbert *et al.* (2013) [418] also aimed to establish a model of feline *T. foetus* adhesion to intestinal epithelial cells. They demonstrated that viable *T. foetus* cells can adhere to monolayers and this adhesion is not reliant on the integrity of the *T. foetus* cytoskeleton. They also compared the cell binding properties of *Pentatrichomonas hominis*, another intestinal trichomonad of cats, and determined that *T. foetus* could adhere in far greater numbers, nearly four times greater in some cases [418].

Singh *et al.* (1999) [414] studied the cytopathic effects of *T. foetus* on bovine vaginal epithelial cells (BVECs) and determined what the role of lipophosphoglycan (LPG) in the adhesion of the parasites. They showed that direct contact of *T. foetus* to the epithelial cells was cytotoxic, the effect being dependent on the number and density of *T. foetus* cells. To investigate the role of LPG, they developed a binding assay. The BVECs were cultured in 24-well plates and allowed to become confluent before the *T. foetus* cells were added. In some wells LPG was added in varying concentrations before the addition of the *T. foetus cells*. The higher the concentration of LPG added prior to infection, the lower the binding of the parasites, suggesting that there is competition for binding sites and showing possible receptor-binding ligands on the cell surface [414].

When HeLa, human epithelial (HEp-2), normal baboon testicular (NBT), and monkey kidney (Vero) cells were exposed to *T. vaginalis* parasites, there was damage to all monolayers [419]. *T. foetus* (strain KV-1) was also found to damage the same monolayers, however, *T. tenax* was found to not induce any cytotoxicity [24]. The actual mechanisms of damage by these trichomonads has been debated for many years [420], however, what is known is that the trichomonads aggregate and it appeared to be the physical attachment of these 'clumps' that caused the damage. Kreiger *et al.* (1985) [420] stated that it is only this physical attachment that causes cell damage, rather than anything being secreted by the parasites themselves.

In contrast, Amin *et al.* (2012) [421] found that when cell-free filtrates of *T. gallinae* were added to a cell monolayer, the monolayer cells had reduced viability. Furthermore, if the same monolayer cells (QT35-Quail Fibroblasts and LMH-chicken liver) were freshly seeded into a new flask and the cell-free filtrate was added, the cells were unable to form confluent monolayers [421]. This indicates that trichomonad secretory products do cause monolayer cell damage, and this is not purely the result of the attachment process [140].

The identity of these secreted factors promoting pathogenesis is not completely understood. The

cysteine protease CP30 has been identified as a virulence factor in *T. foetus* [68] [140]. In 2017 Gould *et al.* [68] showed that CP30 could act as a mediator in the adhesion and cytotoxicity of feline strain *T. foetus* to monolayers. When CP30 was inhibited, the adhesion capabilities and cytotoxicity of *T. foetus* was decreased. Furthermore, expression of CP30 was found in all feline *T. foetus* strains tested [29]. Singh *et al.* (2004) [140] showed that when this protease was inhibited using the inhibitors Nα-p-tosyl-L-lysine chloromethyl ketone (TLCK) and L-trans-epoxysuccinylleucylamide-(4-guanido)-butane (E-64) the adherence of bovine *T. foetus* parasite to bovine vaginal epithelial cells (BVECS) was reduced. However, the same purified cysteine proteases in *T. vaginalis* did not show the same damage to the BVECS, showing some species specificity.

The precise genotype of interacting cells seems to affect the pathology observed. Pindak *et al.* (1986) [422] investigated the effects of *T. vaginalis* on various cell strains to identify those most suitable for cell cytotoxicity experiments. It was seen that, when the *T. vaginalis* cells were inoculated into the cell cultures, only 103 parasites were needed to disrupt the monolayers of HeLa (human), HEp-2 (human) and RK-12 (rabbit kidney). With other cell types, Veros, CHO (chinese hamster ovary) and McCoy (mouse) were tested, 10-100 fold higher levels of T. vaginalis were needed to induce the same amounts of damage [422]. Furthermore, when the *T. vaginalis* cells adhered to all of the monolayers, lesions were visible, showing the monolayers have been damaged.

4.1.4 Pathological Differences Between Trichomonad Strains and Species

While host cell types appear to be affected to varying extents by parasite binding, different trichomonads are known to display interspecific and strain variation in the damage they inflict when placed in the presence of cell monolayers [419]. This damage can depend on the strain and species of trichomonad itself, or the type of monolayer. In a study by Alderete *et al.* (1984) [419], *T. vaginalis* cased twice the damage to HeLa cell monolayers than *T. foetus* and *T. tenax* did not cause any obvious damage [419]. In terms of the monolayers, when *T. vaginalis* was added to several different cell types, clear differences in cytotoxicity were seen [419].

When *T. gallinae* was added to a flask of chicken liver cells and also a flask of fibroblasts, damage to both host cell types was seen, and the cells could not form a cohesive monolayer [421]. Whereas, when *Tetratrichomonas gallinarum* was used, it did not show any visible effect to either cell type. However, the cell filtrates of both trichomonad species were shown to have an effect on the viability of the monolayer cells, suggesting that something secreted was having an impact [421]. This damage occurred both when filtrate was taken from axenic trichomonas culture and infected culture,

ie. the trichomonads had been co-cultured with host cells, and there was less damage when filtrate from *T. gallinarum* was used.

Different phenotypes and subpopulations of *T. vaginalis* were found to have different effects on host cell cytotoxicity [423]. Those *T. vaginalis* cells that did not possess a particular cell surface glycoprotein (MAR) were found to be cytotoxic to HeLa cell monolayers, culminating in their total disruption [423]. Cells that were positive for the glycoprotein had significantly lower levels of cytotoxicity and if the phenotype was altered from positive to negative, the rates of cell monolayer destruction increased accordingly [423]. Significant differences were also seen in surface charge between different strains of *T. vaginalis* and *T. foetus* [424], although the overall mean charge between the two species did not differ. It was found that those strains with the most negative charges were also the most cytotoxic to a mouse monolayer [424].

A proteomic comparison was performed between long-term lab grown and freshly isolated *T. vaginalis* strains to identify if there was a difference in virulence and, if so, which proteins were upregulated in the virulent strain. This analysis identified 29 protein spots that were differentially expressed between the two [171]. Of those 29, 19 were overexpressed in the freshly isolated, more virulent, strain. These included cytoskeletal proteins in addition to metabolic proteins, such as malate dehydrogenase and aldolase [171]. Additionally, it was found that, in the presence of vaginal epithelial cells, the more virulent *T. vaginalis* strain could switch between trophozoite and pseudocyst forms [171] more quickly than the long-term lab grown strain.

Da Rocha-Azevedo *et al.* (2005) [425] compared five clonal subpopulations of *T. foetus* strain K to characterise their cytotoxicity towards epithelial cells. The different subpopulations ablated the monolayers at different rates, ranging from 25-55% although the rates of adhesion and levels of protease activity between them were all similar [425]. The subpopulation that had the highest cytotoxic effects also had high protease activity and those with lower cytotoxic effects had lower activity [425].

Different strains of *T. vaginalis* were also compared to determine whether the length of time grown in a laboratory setting affects their virulence and pathogenicity and which genes and enzymes are upregulated in more pathogenic strains. When two strains of *T. vaginalis*, one a fresh isolate and one long-term lab cultured, were, again, grown in both iron rich and iron depleted media [170] there were clear differences between the strains in terms of growth rate and enzyme activity. The fresh isolate produced four times the enzyme activity (ecto-ATPase) when grown in the iron-rich media than the long-term cultured isolate. In the iron depleted media the fresh isolate again had a

significantly higher enzyme activity, showing clear differences between the strains. These differences could be mirrored in the *T. foetus* stains used within this chapter as one has been grown long term in culture, whereas others are lower passage isolates that have not been cultured long-term within my lab.

As the Belfast strain, which has been used in all prior experiments in this thesis, has been grown in culture for several years, we wanted to identify if there were any changes in terms of cell adhesion or cell damage when compared to other strains. Furthermore, we wanted to examine whether there were significant differences in host-parasite interactions when comparing the new isolates to each other. Additionally, we wanted to find genes that were expressed in the presence of the host cells in all strains as these could be good vaccine candidates.

4.1.5 Aims and Objectives

The aim of this chapter was to produce RNA-Seq data for *T. foetus* parasites in the presence of a host cell monolayer, in this case Madine-Darby Canine Kidney cells (MDCKs) and to identify the differentially expressed genes. In order to achieve this we aimed to:

1) Test several strains of *T. foetus* and *T. mobilensis* for cytotoxicity against an MDCK monolayer
2) Test several strains of *T. foetus* and *T. mobilensis* for adherence to an MDCK monolayer
3) Produce transcriptomes from *T. foetus* cells in the presence of an MDCK monolayer
4) Identify differentially expressed genes between *T. foetus* cells grown axenically and *T. foetus* that had been incubated in the presence of an MDCK monolayer.

4.2 Methods

4.2.1 MDCK cell culture

Madine-Darby Canine Kidney (MDCK) epithelial cells were grown in DMEM with 10% foetal bovine serum (FBS) and 1% Penicillin/Streptomycin at 37°C with 5% $C0_2$. This cell type was chosen because it has been previously used to examine trichomonad-host cell interactions [426] [29].

4.2.2 *T. foetus* Strain and Cell Culture

Multiple strains of *T. foetus*, along with a related parasite, *T. mobilensis* were used in the MDCK cell monolayer experiments. This was to identify whether all strains and species showed the same cell adherence and cytotoxicity, particularly as the Belfast strain had been grown in a laboratory setting for many passages and this may have affected the virulence. In total, four other *T. foetus* strains were used along with *T. mobilensis*:

1) *T. foetus* KV-1 strain-naturally infected from Czechoslovakia (1962) ATCC 30924
2) *T. foetus* BP-4 preputial washings from Beltsville Maryland (1956) ATCC 3003
3) *T. foetus* UT Preputial washings from Beltsville Maryland (1967) ATCC 30232
4) *T. foetus* DK-2 Preputial washings from Davis California (1967) ATCC 30231
5) *T. mobilensis* USA-M776 Bolivian squirrel monkey (1984) ATCC 50116

Cells were initially grown in ATCC entamoeba medium before being passaged into Diamond media. All cells were incubated at 35°C.

4.2.3 *T. foetus* Growth Curves

The five strains of *T. foetus*: Belfast, KV-1, BP-4, UT and DK-2 and *T. mobilensis* were each grown in 15ml tubes of Diamond media. 1ml of cells from these 15ml growth tubes, at 24 hours post passage, was spun down at 1000rpm for 5 minutes. The pellet produced was resuspended in fresh Diamond media to give a cell count of approximately $1x10^6$ cells for each strain. 200μl of this cell suspension was added to a 15ml falcon tube containing fresh Diamond media. 10μl samples were taken at different intervals: 30 minutes, 1 hour, 2 hours, 6 hours, 12 hours, 24 hours and 48 hours. Growth curves were then created for each strain.

Growth curves were also produced when the *T. foetus* cells were cultured in DMEM in the presence

of MDCK cells. Briefly, 1ml of cells from 15ml growth tubes was spun down at 1000rpm for 5 minutes. The pellet was resuspended in fresh media to give a cell count of approximately 1x10^5 cells. 200μl of each cell suspension was added to a well in a 46 well plate, already containing 80% confluent MDCK cells. The plate was incubated at 35°C and 10μl samples were taken at different time intervals stated previously for the initial growth curves.

4.2.4 MDCK Cell Binding Assays

MDCK cells were grown to confluency to approximately 2x10^6 cells in a T25 flask. *T. foetus* cells were grown in Diamond media to approximately 5x10^6 cells.

The *T. foetus* cells were spun down 1000rpm for 5 minutes and washed in PBS. They were then added at a ratio of 10:1 *T.foetus*:MDCK cells. Initially the cells were incubated with the parasites from 30 minutes to 48 hours and samples were taken at different time points: 30 minutes, 1 hour, 2 hours, 6 hours, 12 hours, 24 hours and 48 hours. The media was removed and cells washed gently five times with PBS. The T25 flask was then imaged under an inverted light microscope. The percentage of MDCK cells with at least one *T. foetus* cell attached was calculated.

4.2.5 MDCK Cell Cytotoxicity Assays

In the same way as the binding assay, *T. foetus* cells were added at a ratio of 10:1. The media was removed and saved and 100μl trypsin-EDTA was added.
After 2 minutes, when the cells were released and the old media was added to neutralise the trypsin. 100μl of the sample was taken and added to 100μl trypan blue. This was then placed in a haemocytometer so cell counts could be taken and dead cells could be identified. Any cells that had been stained with the trypan blue were assumed to be dead. The haemocytometer was then imaged using an inverted light microscope.

4.2.6 Effect of Cell Number on Cell Binding and Cytotoxicity

In order to determine how much of an impact the initial cell number has on the numbers of cells binding, two different concentrations of cells were added to a 6-well plate, with each well containing an MDCK cell monolayer.
200μl of approximately 1x10^5 cells were added to wells (1-3 across).
20μl of approximately 1x10^5 cells were added to wells (4-6 across).

The supernatant was removed and the cells washed with PBS as previously stated. The number of MDCK cells with at least one *T. foetus* cell attached per field of view were counted.

4.2.7 The Effect of Cell Growth Stages on MDCK Cytotoxicity and Cell Death

In all other experiments, the cells were taken 24 hours after passage, (the middle of the log growth phase). Trials were done using cells 12 hours after passage (early log phase) to see if and/or how this affected the MDCK cell death and adhesion. This was to see whether cells that are in earlier stages of growth possess any differences in pathogenicity than later growth stage cells. The trial was set up as stated previously in the cell cytotoxicity assays with *T. foetus* cells from early-log growth phase added to three wells of a 6-well plate containing a confluent MDCK monolayer and mid-log growth phase *T. foetus* cells added to the remaining three. Another plate was used as cell death controls, containing confluent MDCK cell monolayers that had been originally passaged on the same day as the experimental samples.

The Effect of Cell Growth Stages on MDCK Cytotoxicity-sub 2 hours

To further identify whether there are differences in early infection rather than later infection and to find the optimum points for harvesting cells for transcriptomics, and to determine whether early growth phase *T. foetus* cells show greater cell adhesion and cytotoxicity assays were performed using both, 24 hour and 12 hour post-passage *T. foetus* cells at 15 minute time points up to 2 hours. The time points were: 15 minutes (post addition of *T. foetus* cells), 30 minutes, 45 minutes, 1 hour, 1 hour 15 minutes, 1 hour 30 minutes, 1 hour 45 minutes and 2 hours.

4.2.8 The Effect of Trichomonad Passage Number on MDCK Cell Death

The aim was to see if the passage number of the trichomonad cells had an effect on the cell death rate of the MDCKs, i.e. if they became more or less damaging the longer they have been allowed to multiply and have been passaged. When trichomonads are kept in cell culture for long periods of time and several passages, their phenotypes can be changed, leading to reduced cell cytotoxicity or attachment. The use of early passages to was to diminish this effect as far as possible so that effects seen were due to strain differences rather than lab culture conditions. To test, early passage numbers (passage 1-3 of 'new strains' and passage 10 of Belfast) were used in adherence and trypan

blue exclusion assays (Table 4.1). The results were then compared to 'late' passage samples (passage 15 or above for all strains).

Strain	'Early' Passage Number	'Late' Passage Number
Belfast	10	35
BP4	1	15
UT	1	18
KV1	2	15
T.mobilensis	3	19
DK2	1	20

Table 4.1: Passage number of trichomonads used in MDCK cytotoxicity experiments. Both 'Early' (passage 10 and below) and 'Late' (over passage 15) passages of the trichomonads were added to a confluent MDCK monolayer for up to 48 hours. A trypan blue assay was performed to test for cell death.

4.2.9 Supernatant Effect on Cell Death

In order to test whether it is the attachment of the trichomonads physically that is damaging the MDCK cell monolayer or whether it is something being secreted by the trichomonads, trials were set up using supernatant only to test for cell damage. Supernatant from one well of a 6-well plate that had contained trichomonads and an MDCK monolayer was removed and added to a well containing an MDCK monolayer only. These MDCKs had never been in contact with any trichomonads. 'Early' and 'Late' supernatant was used for the trials. In each case 'early' supernatant refers to supernatant from the well where trichomonad cells have be in contact with the MDCK monolayer for 1 hour and 'late' refers to the supernatant from a well were the trichomonas cells have been in contact with an MDCK monolayer for 24 hours.

Four different trials were performed:
1)'Early' supernatant added to MDCKs for 1 hour.
2)'Early' supernatant added to MDCKs for 24 hours.
3)'Late' supernatant added to MDCKs for 1 hour.
4)'Late' supernatant added to MDCKs for 24 hours.

Initial cell counts from the 15ml growth tubes were taken. 1ml from each was taken, spun down at 1200rpm for 5 minutes and resuspended to give $4x10^4$ cells.
200μl of this suspension of trichomonad cells was added to the MDCK wells initially. The plate was incubated at 35°C. After the specified time, either 1 hour or 24 hours, the media was aspirated and spun down at 1200rpm to remove the trichomonads. 200μl of supernatant was removed from the

new MDCK well before the cell suspension or control suspensions were added. The supernatant from the trichomonad infected well, was then added to fresh MDCK cells for the specified time (1 hour or 24 hours).

After this time 100μl of the media was added to 100μl trypan blue. 10μl was then imaged using an inverted microscope (10x magnification) and percentage cell death was calculated. The early and late supernatant samples were compared, as was the 1 hour and 24 hour time points. There were two controls used which were:

1) PBS only added to the MDCK wells

2) Diamond media only added to the MDCK wells.

4.2.10 Cell Detachment

The aim was to see how many trichomonad cells can reasonably be expected to detached from the cell monolayer and thus determine how many cells need to be added initially to get the optimal number for successful RNA extractions.

200μl of approximately 1x10^5 cells were added to wells (1-3 across).

20μl of approximately 1x10^5 cells were added to wells (4-6 across).

The media was then aspirated. 200μl of trypsin-EDTA was added to each well and the plate was left for 3 minutes at 35°C. When cells appeared to be rounded and detaching (visualised under 10x magnification using an inverted microscope) 200μl of DMEM, containing 10% FBS, was added to neutralise the trypsin-EDTA.

10μl was taken from the neutralised solution and imaged using an inverted microscope. The procedure was also repeated using cells that had been washed five time with PBS to see how many 'attached' cells would remain within the flask. This was to gauge how many cells there were likely to be for starting material for the RNA extractions.

4.2.11 Separation of cells

In order to get the optimal *T. foetus* transcriptomes, a pure *T. foetus* after the co-incubation with MDCK cells would be ideal. In order to achieve this, a way had to be devised to separate both cell types. After co-incubation of trichomonads and MDCK, the monolayer was washed five times with PBS, trypsin-EDTA was added for 30 seconds. This allowed the trichomonads to detach from the MDCKs whilst minimising the number of MDCKs detaching from the flask.

4.2.12 RNA extraction

Initially, both the Belfast and DK2 strains were selected for the experiment; the former because it has been used in all other proteomics trials, the latter because it gave the highest percentage of cell damage and, therefore, may be the most 'pathogenic'. Given that the dynamics of gene expression when parasites initially contact host epithelia are not know, three different time points were chosen for sampling (30 minutes, 1 hour and 24 hours), in order to accommodate distinct effects at different times post contact.

Three technical replicates and three biological replicates were taken for each time point and each strain. For each time point three flasks were set up and the RNA extracted. This was then repeated three separate times to increase replication. Due to time and reagent constraints it was decided that samples would only be taken from DK2 cells as these produced a stronger cytotoxic response (Figures 4.4 and 4.8). For the transcriptomic DK2 controls, cells were grown in a T25 cell culture flask for 24 hours and then all cells were harvested and the RNA was extracted as stated in Chapter 2.

4.2.13 RNA preparation and sequencing

Due to issues of quality with certain samples (Table 4.2), not all library preparations could be produced using poly-A selection, as was the case in Chapter 1, hence, the lower quality samples were processed using the 'Zymo' (Zymo Research) isolation kit. This kit was able to process lower levels of RNA and convert the RNA to cDNA before a universal depletion to produce the libraries.

All RNA samples were sequenced on the Illumina Novaseq. The reads produced (ENA Study PRJB39464) were processed in the same way as Chapter 1 using Cutadapt v1.2.1 [300] and Sickle v1.200 [301]. The resulting fastq files were uploaded to the Galaxy server [302] and mapped to the completed Belfast genome assembly sequence using the program Hisat2 [303] [304].

Sample Number	Experiment	Library Preparation
1	Trophozoite Trypsin Wash (30 min) 1	Poly-A Selection
2	Trophozoite Trypsin Wash (30 min) 2	Poly-A Selection
3	Trophozoite Trypsin Wash (30 min) 3	Poly-A Selection
4	Pseudocyst Trypsin Wash (30 min) 1	Zymo-kit
5	Pseudocyst Trypsin Wash (30 min) 2	Zymo-kit
6	Pseudocyst Trypsin Wash (30 min) 3	Zymo-kit
7	Trophozoite Trypsin Wash (1 hour) 1	Poly-A Selection
8	Trophozoite Trypsin Wash (1 hour) 2	Poly-A Selection
9	Trophozoite Trypsin Wash (1 hour) 3	Zymo-kit
10	Pseudocyst Trypsin Wash (1 hour) 1	Zymo-kit
11	Pseudocyst Trypsin Wash (1 hour) 2	Zymo-kit
12	Pseudocyst Trypsin Wash (1 hour) 3	Poly-A Selection
13	Trophozoite Trypsin Wash (24 hours) 1	Zymo-kit
14	Trophozoite Trypsin Wash (24 hours) 2	Zymo-kit
15	Trophozoite Trypsin Wash (24 hours) 3	Zymo-kit
16	Pseudocyst Trypsin Wash (24 hours) 1	Zymo-kit
17	Pseudocyst Trypsin Wash (24 hours) 2	Zymo-kit
18	Pseudocyst Trypsin Wash (24 hours) 3	Zymo-kit
19	Trophozoite PBS Wash (1 hour)	Zymo-kit
20	Trophozoite PBS Wash (24 hours)	Poly-A Selection
21	Pseudocyst PBS Wash (24 hours	Poly-A Selection
22	MDCK cells only	Zymo-kit
23	DK2 Control 1	Zymo-kit
24	DK2 Control 2	Zymo-kit
25	DK2 Control 3	Poly-A Selection
26	DK2 Supernatant 1	Poly-A Selection
27	DK2 Supernatant 2	Poly-A Selection
28	DK2 Supernatant 3	Poly-A Selection

Table 4.2: The methods of library preparation for host-monolayer RNA-Seq samples. Where some samples had lower quality, the 'Zymo' kit was used rather than Poly-A selection.

4.3 Results

4.3.1 Trichomonad cell growth

When cells were grown in Diamond media they were passaged at mid log phase and samples were taken at 1h, 2h, 6h, 12h, 24h, 48h and 72h and cell counts made. All strains show the same shaped curve, with the cell numbers increasing from the initial passage, peaking at 48 hours and dropping off post 48 hours (Figure 4.1). This may be due to a lack of space and nutrients and a build-up of toxic waste products. The Belfast strain had the highest total cell count of all the strains, reaching over 350x10^4 cells per ml, potentially due to the fact that it had been grown in cell culture for many passages and was well adapted. All samples showed a lag phase up to 6 hours, with *T. mobilensis* having a lag phase of 12 hours. At their peak three samples: Belfast, *T. mobilensis* and UT had a density of over 300x10^4.

When the *T. foetus* cells were grown with MDCK cells a similar growth pattern was seen with all strains reaching a peak before decreasing (Figure 4.2). However, the peak of the cell numbers for all strains was at 24 hours rather than 48 hours and the decrease in cell numbers occurs more slowly. It may be due to the increase in nutrients available from the monolayer. Furthermore, the peak number of cells when grown in the T25 flask in the presence of a monolayer was lower for all samples than the *T. foetus* grown alone, with peak numbers reaching between 75 and 125x10^4 cells per ml compared to over 250x10^4 cells.

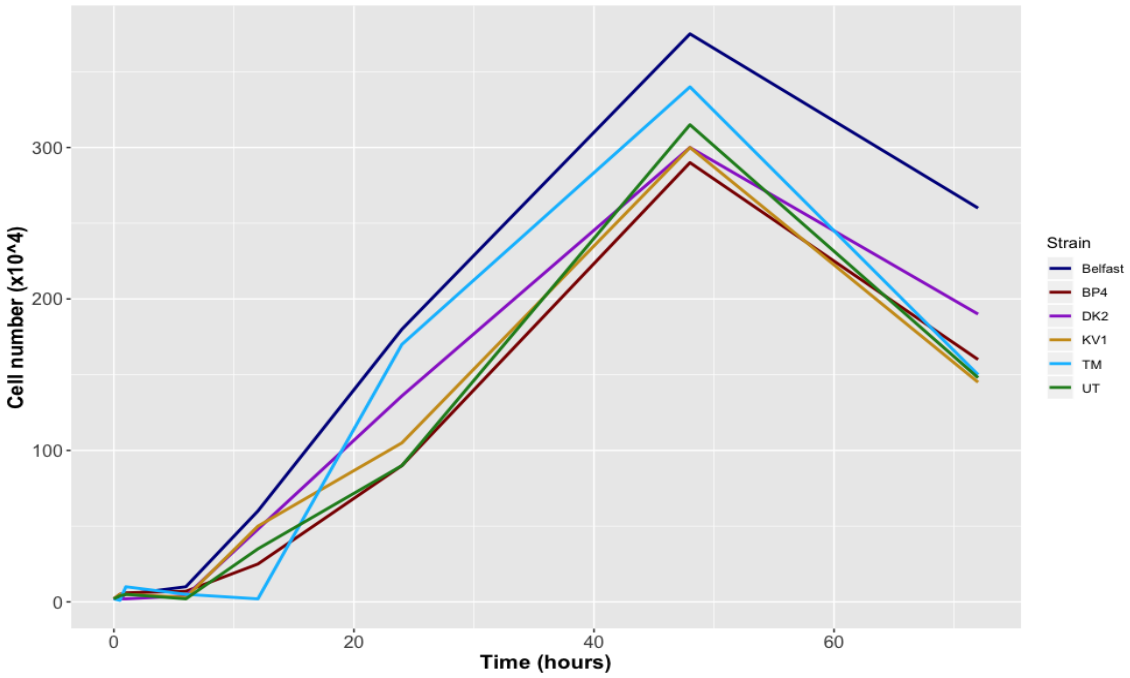

Figure 4.1: Growth Curves of five strains of *T. foetus*: Belfast, BP-4 DK-2, KV-1 and UT and *T. mobilensis* (TM) over 72 hours. All cells were grown in 15ml tubes of Diamond media at 37°C. Cell numbers were calculated using a haemocytometer.

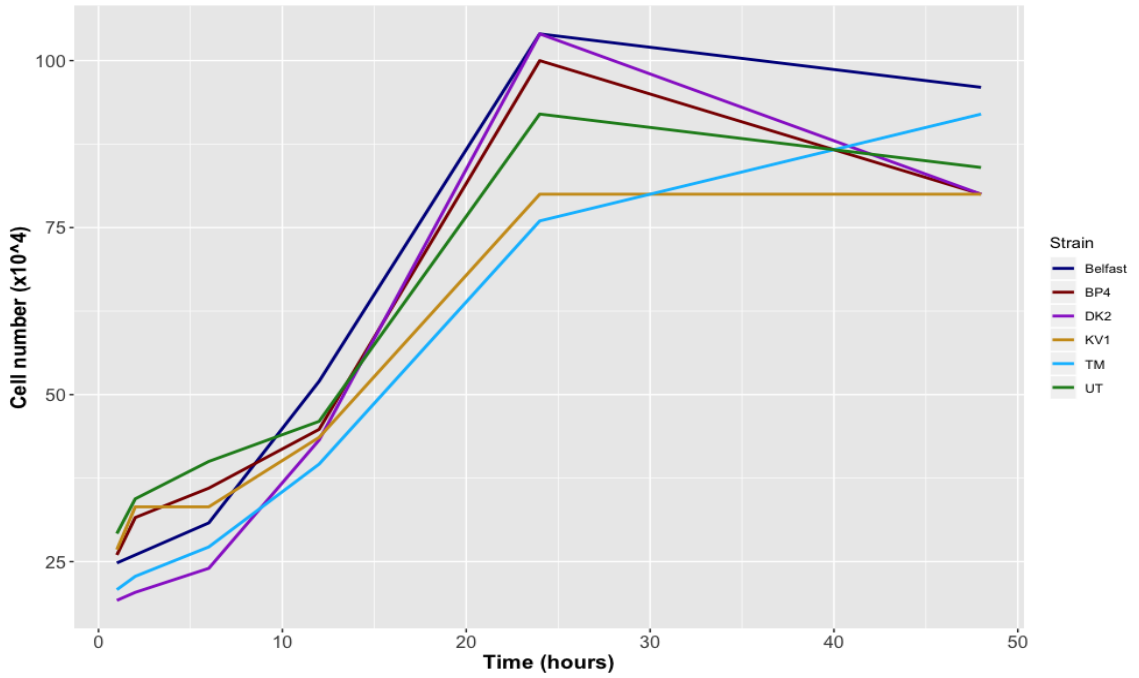

Figure 4.2: Growth Curves of five strains of *T. foetus*: Belfast, BP-4 DK-2, KV-1 and UT and *T. mobilensis* (TM) over 48 hours. All cells were grown in T25 cell culture flasks of DMEM containing a confluent cell monolayer of MDCK cells at 37°C. Cell numbers were calculated using a haemocytometer.

4.3.2 Binding assays

As time went on the number of *T. foetus* cells bound to the MDCK monolayer increased (Figure 4.3). This was particularly marked over the first six hours, with the DK2 strains showing much higher adhesion, with over 50 cells bound per field of veiw, than all other strains. This reached 250 cells bound per field of view after 48 hours for the DK2 strain, with the *T. mobilensis* strain being the next highest with over 150 cells bound. All other strains had less than 100 cells bound at 48 hours.

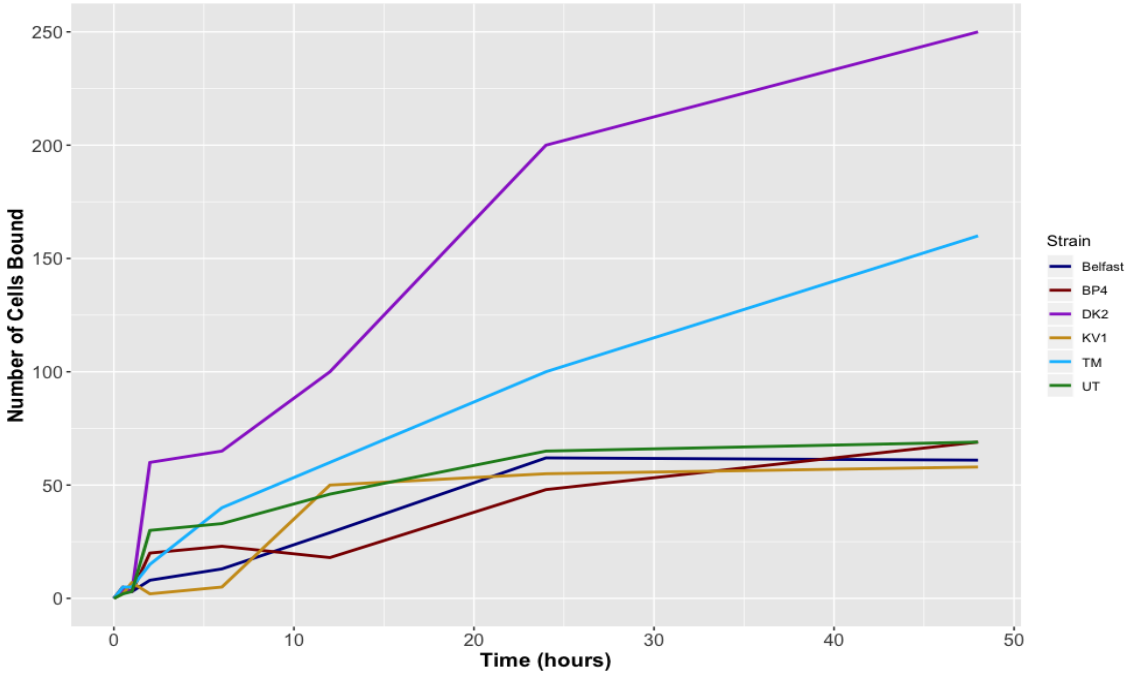

Figure 4.3: Cell adhesion of five strains of *T. foetus*: Belfast, BP-4 DK-2, KV-1 and UT and *T. mobilensis* to a confluent MDCK monolayer over 48 hours. Cells were counted as bound if at least one parasite cell attached to it after five PBS wash steps.

4.3.3 Cytotoxicity Assays

As time increased the proportion of dead MDCKs cells from the sample of supernatant increased. The most dramatic increase was in the first hour (Figure 4.4) but the percentage cell death increased after this point up to 24 hours, albeit more slowly. The DK2 strain appeared to be the most damaging or cytotoxic when compared to the other strains, with the highest percentage cell death reaching over 70%. The UT strain also reached a high percentage cell death at over 65%. The Belfast strain, conversely, had a relatively low cytotoxic or destructive capability as the death rate did not reach 20%, even at 24 hours post addition to the monolayer

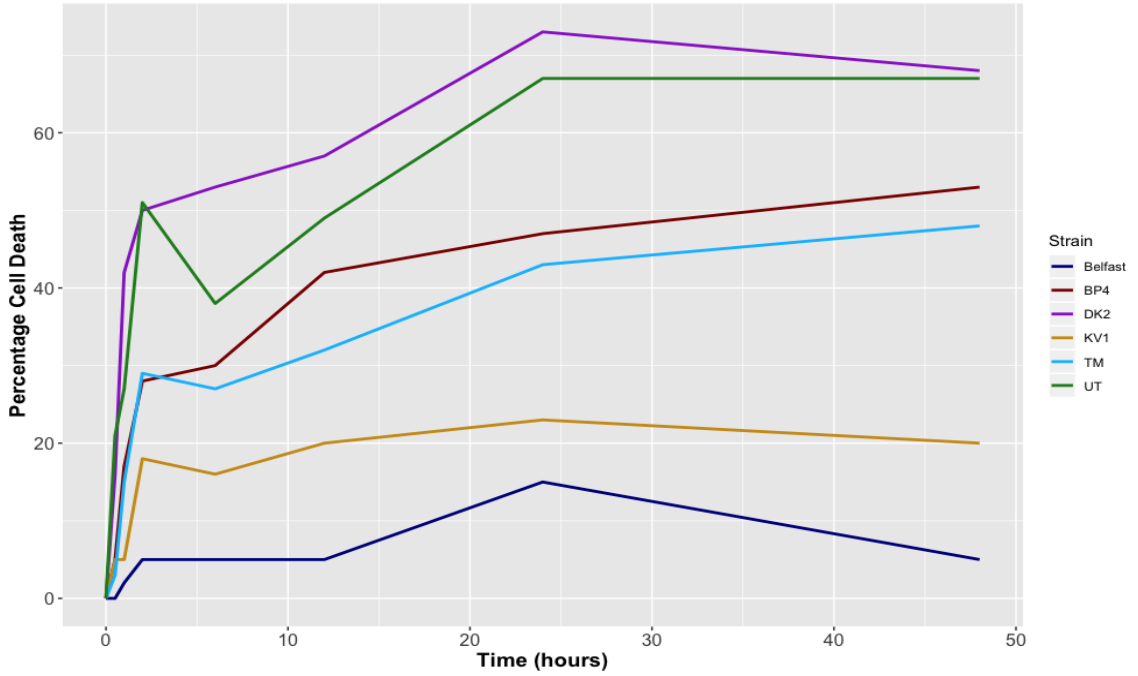

Figure 4.4: Cell death of MDCK cells when five strains of *T. foetus* trophozoites: Belfast, BP-4 DK-2, KV-1 and UT and *T. mobilensis* were added for up to 48 hours. Parasite trophozoites in log growth phase were added to a confluent monolayer of MDCK cells. Cell death was determined by way of a trypan blue exclusion assay.

4.3.4 The Effect of Cell Growth Stages on MDCK Cell Death

T. foetus cells from axenic culture, were taken at different points in their growth cycles; 12 hours and 24 hours after passaging, corresponding with the beginning and middle of log phase respectively. This was in order to determine whether the *T. foetus* cells have different phenotypes with regards to adhesion and cytotoxicity depending on their growth phase. The percentage cell death of the MDCK cells were measured for 2 hours, with samples being taken every 15 minutes. For the early log phase samples (12 hours post original passage) the percentage cell death increased from 0 to 2 hours in all strains (Figure 4.5). The same is also true of the strains 24 hours post passage (Figure 4.6) and in both cases it was, again, the strain DK2 that seems to produce the most marked effect in terms of cell death, increasing to 50% in the 12 hour samples and to 45% in the 24 hour samples. Cell death induced by the Belfast strain was very low in both trials, never reaching more than 15% and in some cases was lower than the control samples of MDCKs with no trichomonads suggesting there is very little, if any, destructive capability. Overall, the growth phase of the *T. foetus* cells did not have an effect on the cytotoxicity of MDCKs.

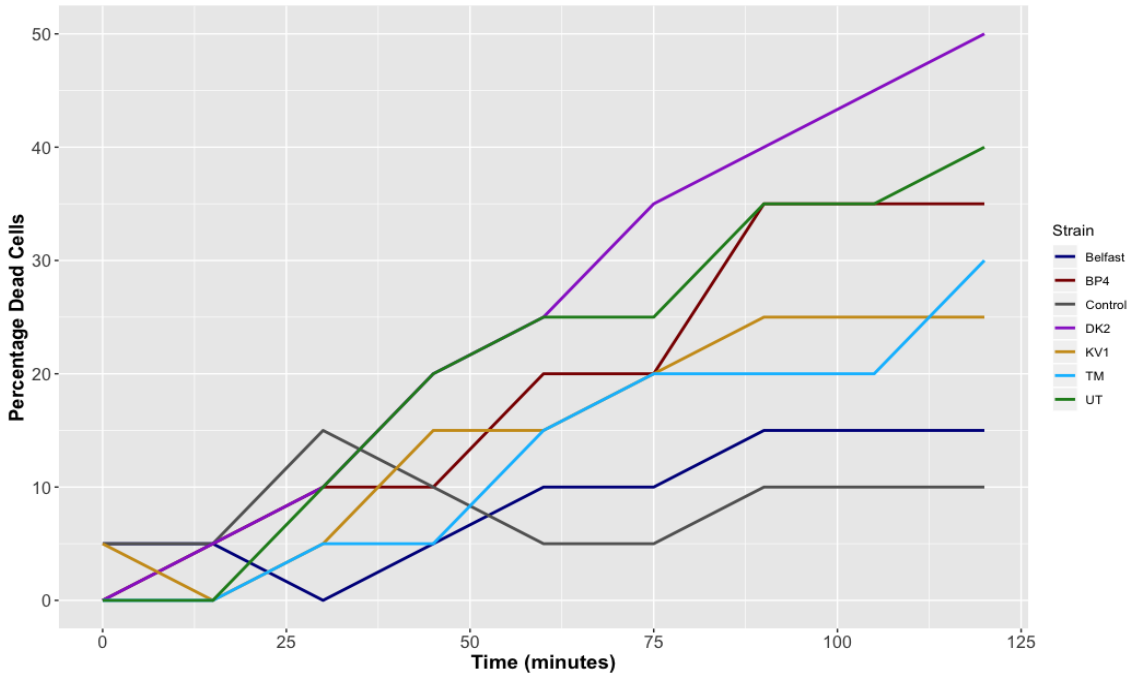

Figure 4.5: Cell death of MDCK cells when five strains of *T. foetus*: Belfast, BP-4 DK-2, KV-1 and UT and *T. mobilensis* were added for up to 2 hours with samples taken every 15 minutes. The *T. foetus* cells had been growing for 12 hours post passage which corresponds with the beginning of log phase growth. Cell death was determined using a trypan blue exclusion assay.

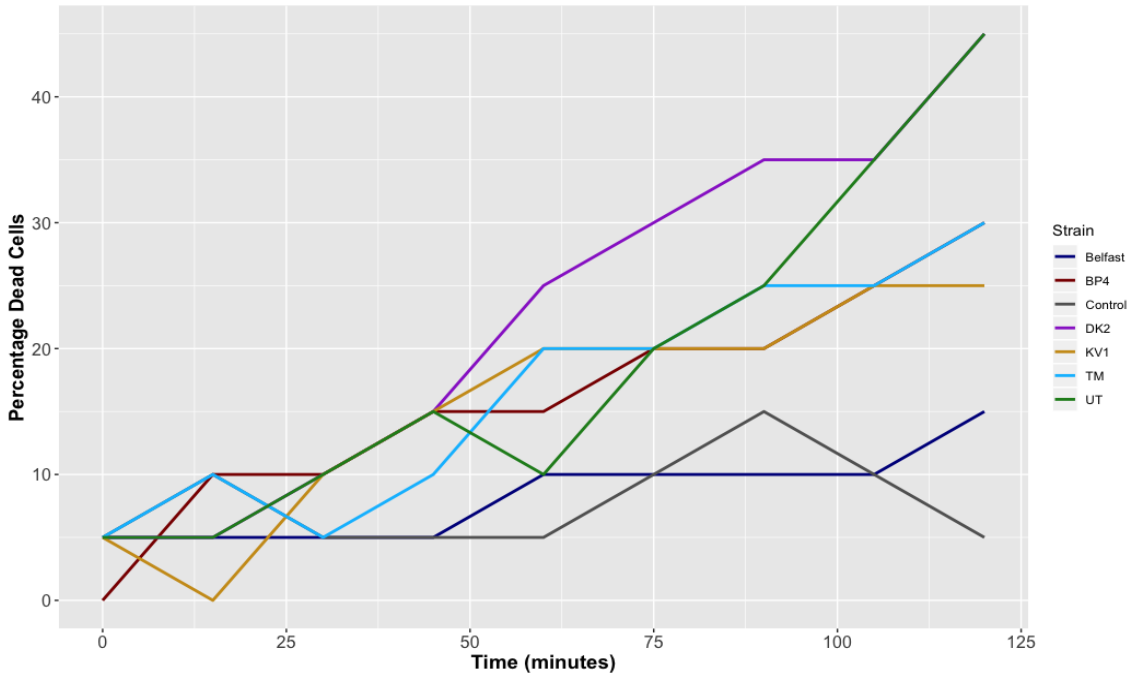

Figure 4.6: Cell death of MDCK cells when five strains of *T. foetus*: Belfast, BP-4 DK-2, KV-1 and UT and *T. mobilensis* were added for up to 2 hours with samples taken every 15 minutes. The *T. foetus* cells had been growing for 24 hours post passage which corresponds with the middle of log phase growth. Cell death was determined using a trypan blue exclusion assay.

4.3.5 Effect of Trichomonad Passage Number on MDCK Cell Death

'Early' passage number *T. foetus* cells (passage 10 or less) were also used to examine cytotoxicity and cell damage towards MDCK cells. This was to determine whether the effects seen between strains were due purely to biological differences or whether their length of time in culture also has an effect. The early passage *T. foetus* cells did not appear to have a different effect than using older passages (Figure 4.4). The cell death percentages were comparable with the higher passages (Figure 4.7) as four of the six strains (DK-2, BP-4, UT and *T. mobilensis*) still reached over 40% cell death, whilst the other two strains, KV-1 and Belfast induced less than 20% cell death. The same strains, DK-2 and UT were the most cytotoxic and damaging, as was seen in the original experiment with the late passages. Overall, the passage number did not have an effect on the cytotoxicity of the different *T. foetus* strains towards MDCK cells.

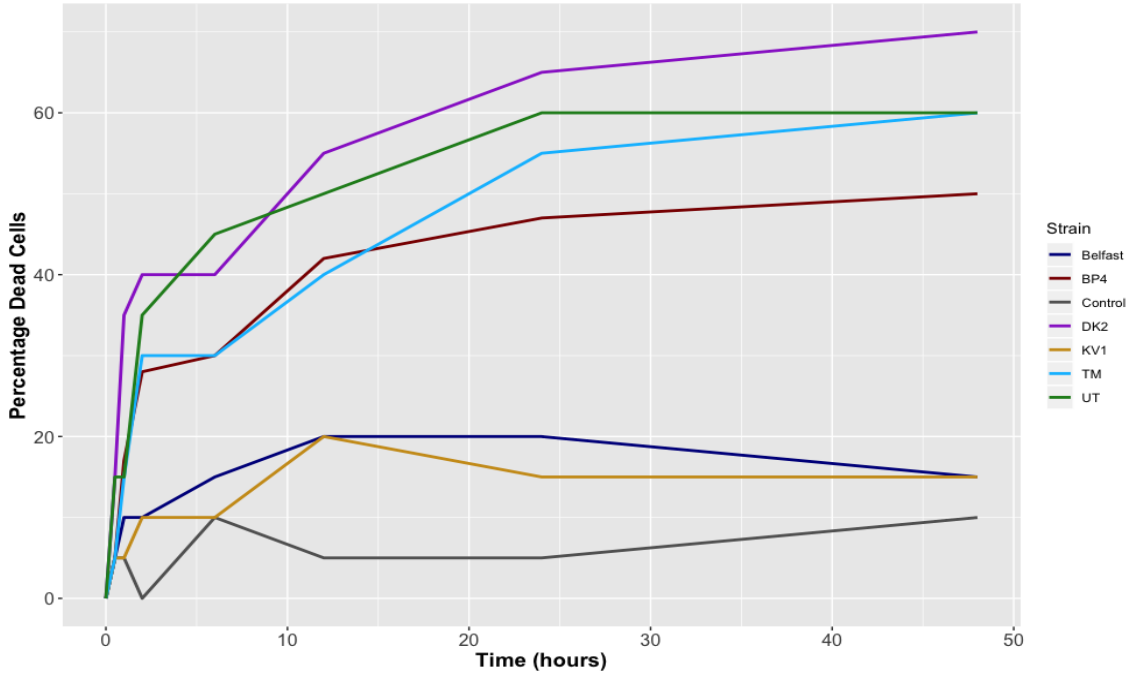

Figure 4.7: Cell death of a confluent monolayer of MDCK cells when five strains of *T. foetus*: Belfast, BP-4 DK-2, KV-1 and UT and *T. mobilensis* were added for up to 48h. All *T. foetus* strains were early passage numbers, P10 or below. Cell death was determined using a trypan blue exclusion assay.

4.3.6 Secretome

Initially supernatant from growing trichomonads was added to a confluent MDCK monolayer to see if there was any effect and whether it was likely that there are secretions from the parasite causing the cell monolayer damage (Figure 4.8). This showed that there was still damage to cells occurring, even without the presence of the physical trichomonad cell bodies. The percentage cell death also followed the same pattern as the initial trichomonad cytotoxicity experiments (Figures 4.4, 4.5 and 4.6) with DK2 showing a far higher effect on the MDCK mortality, with a cell death of over 70% at 24 hours. Similarly, the Belfast strain shows the lowest cell death induction compared to the other strains, never reaching 20%.

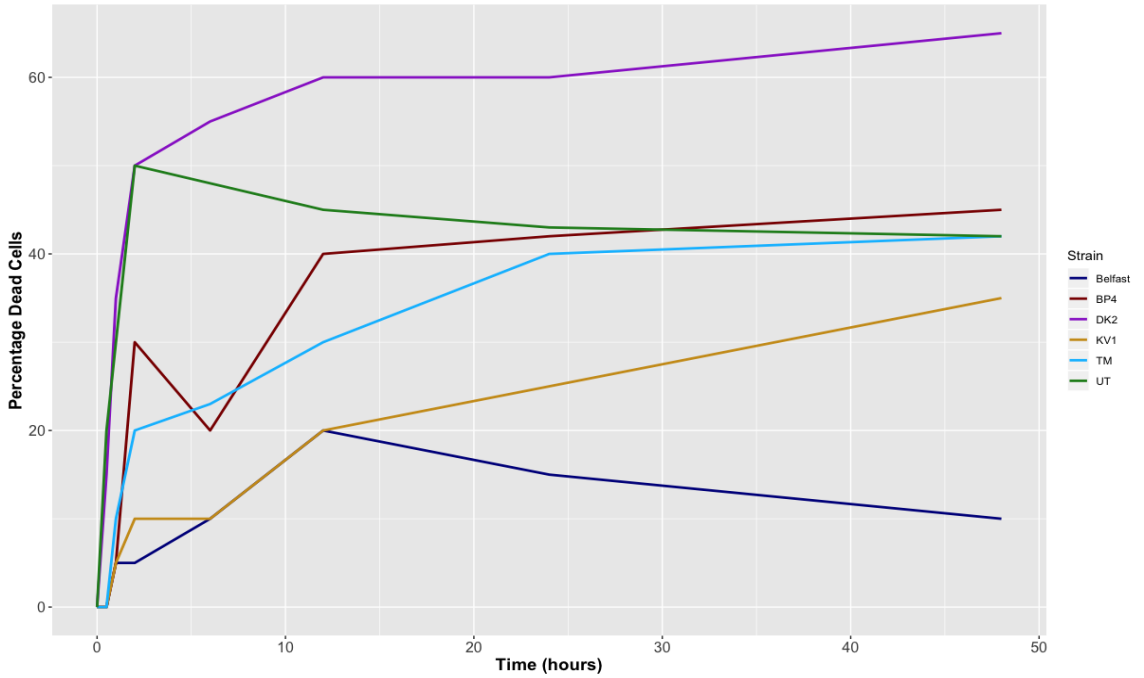

Figure 4.8: Cell death of MDCK Monolayers when *T. foetus* supernatant was added. *T. foetus* cells were grown in diamond media to mid-log phase. The cells were spun down and removed, leaving the supernatant containing any cellular secretions. This supernatant was then added to a confluent MDCK monolayer and the percentage cell death calculated using a trypan blue exclusion assay.

When supernatant from trichomonads that had been growing for 1 hour were added to an MDCK monolayer for 1 hour (Figure 4.9), the UT and DK-2 strains showed the highest levels of cell death, reaching 60%. BP-4 and *T. mobilensis* also have a high level of cell death at 50% and KV-1 and Belfast strains are lower at 30% and 20% respectively. This mirrors the results seen in the other cytotoxicity experiments.

The rates of cell death for each strain were compared between the whole cell sample and supernatant sample by way of a paired t-test, there was no significant difference seen for any of the strains.

When supernatant from trichomonads that had been growing for 1 hour were added to an MDCK monolayer for 24 hours the results are largely the same as leaving the supernatant for 1 hour (Figure 4.9), with DK2 and UT showing high levels of cell death (60%) and Belfast showing lower levels (20%). However, the level of cell death when the KV-1 strain is added to the monolayer is much higher at 24 hours than at 1 hour, reaching 50% rather than 30%.

When supernatant from trichomonads that had been growing for 24 hours were added to an MDCK

monolayer for 1 hour the results were similar, in that strain DK2 had the highest death percentage(Figure 4.10). In this case it was even higher than the 'early' supernatant at 80% cell death verses 60%. Belfast and the controls also showed higher death rates than in the 'early' supernatant trials and, interestingly, KV-1 was much lower at only 20% cell death rate compared to all other trials when the death rate ranged between 30-60% and was higher than the controls.

When supernatant from trichomonads that had been growing for 24 hours were added to an MDCK monolayer for 24 hours (Figure 4.10) the results, again, were broadly the same. With DK2 showing the highest rates of cell death (80%), whereas the Belfast strain induced relatively low levels (40%) comparatively. The KV-1 strain appeared to need longer periods of time to induce cell damage as, when the supernatant was added to the MDCK cells for 24 hours, the death rate was 50% or over in both cases. However, when the supernatant was added for only one hour, the cell death rates were much lower, at 20% and 30%.

As in almost all cases, apart from the 24 hour KV-1 sample added to a monolayer for 1 hour, the percentage cell death was always higher in cases where trichomonad supernatant was used opposed to the controls. This directly contradicts Krieger *et al.* (1985) findings as they stated that it was direct contact alone that damaged the CHO cell monolayers, rather than anything that is being secreted by the trichomonad. However, these results are in keeping with the findings of Amin *et al.* (2012) [421] as they found evidence of secreted proteins, such as cysteine peptidases, that could damage the monolayer without direct cell-cell contact.

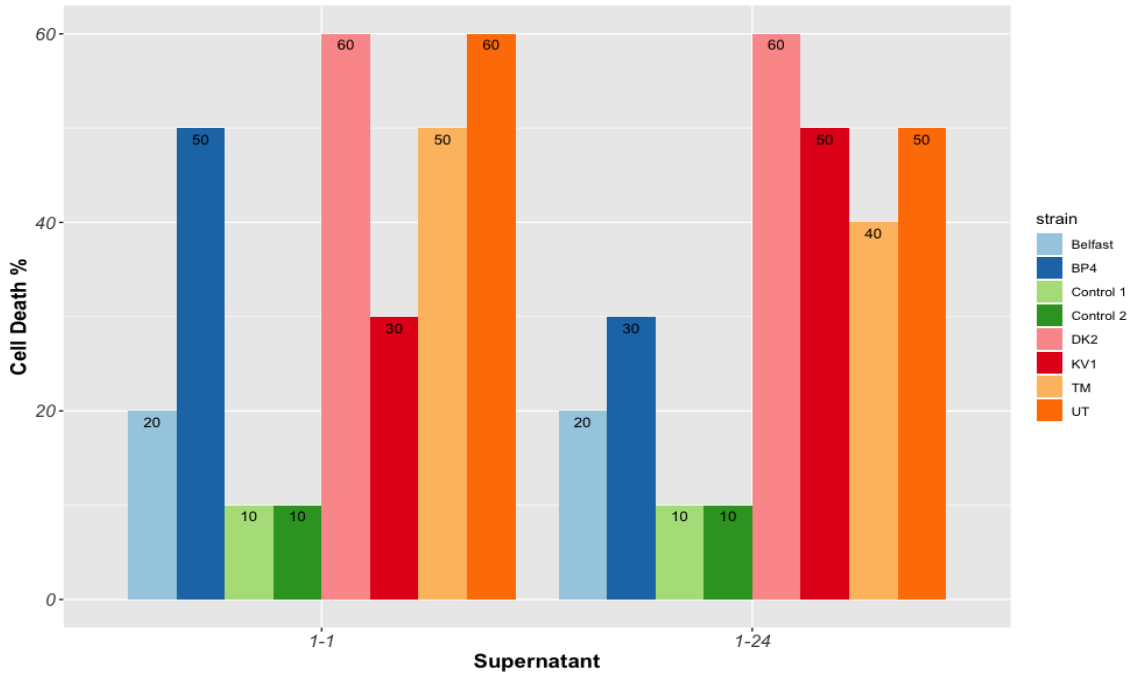

Figure 4.9: MDCK cell death when trichomonad supernatant is added for 1 hour. The supernatant had been extracted from four strains of *T. foetus*: Belfast, BP-4, DK-2, KV1 and UT and *T. mobilensis*. Two controls were also used consisting of 1) PBS and 2) Diamond media only. The supernatant was obtained from trichomonads that had been growing in the presence of an MDCK monolayer for one hour, this supernatant was then added to a fresh MDCK monolayer for 1 hour (1-1) or 24 hours (1-24) respectively.

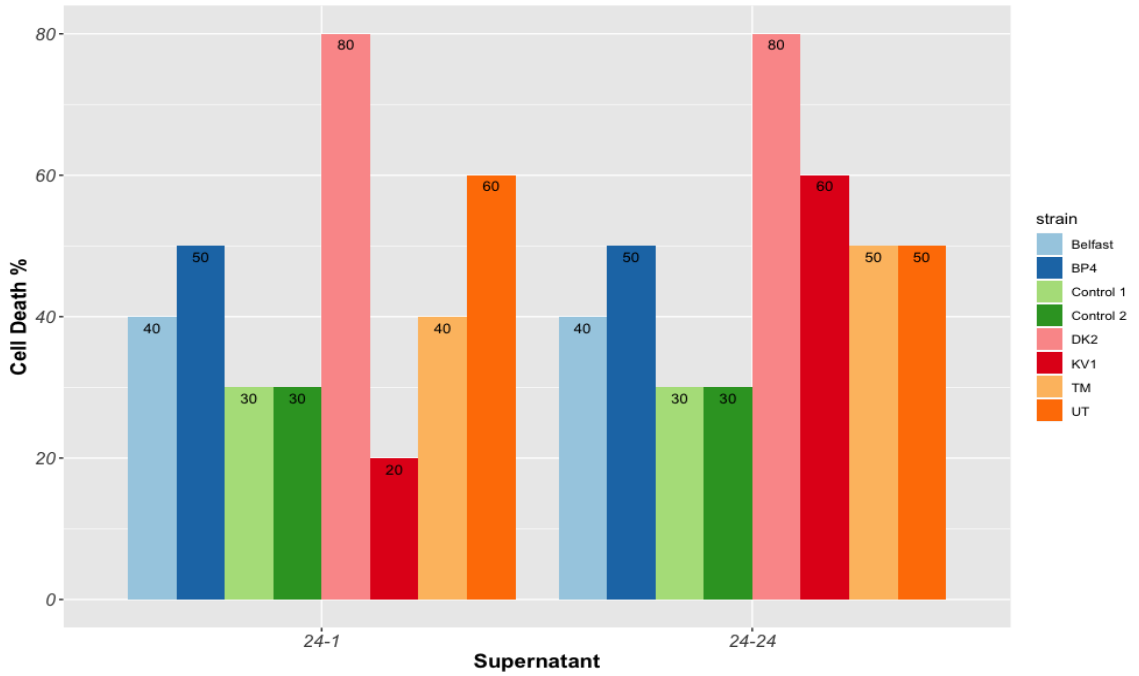

Figure 4.10: MDCK cell death when trichomonad supernatant is added for 1 hour. The supernatant had been extracted from four strains of *T. foetus*: Belfast, BP-4, DK-2, KV1 and UT and *T. mobilensis*. Two controls were also used consisting of 1) PBS and 2) Diamond media only. The supernatant was obtained from trichomonads that had been growing in the presence of an MDCK monolayer for 24 hours, this supernatant was then added to a fresh MDCK monolayer for 1 hour (24-1) or 24 hours (24-24) respectively.

4.3.7 Cell Detachment

To see how many trichomonad cells were needed to be added initially to the monolayer in order to obtain successful RNA extractions. Trypsin was added to the MDCK-trichomonad co-culture and the number of *T foetus* cells that detached after three minutes was measured ('unwashed' samples) (Table4.3). This was also repeated for monolayers that had been washed using PBS prior to trypsinisation ('washed' samples) in order to obtain only those *T. foetus* cells that had been physically attached to the monolayer. Overall the number of cells retrieved after the monolayer had been washed was 3-10 times fewer than the unwashed samples.

Strain	Cell counts (unwashed)	Cell counts (washed)
Belfast	$14x10^4$	$1x10^4$
BP4	$9.5x10^4$	$1x10^4$
UT	$12.5x10^4$	$3.5x10^4$
KV1	$26x10^4$	$7.5x10^4$
T.mob	$19.5x10^4$	$9.5x10^4$
DK2	$21.5x10^4$	$3x10^4$

Table 4.3: Cell counts per well of 'unwashed' and 'washed' cells from one well of a 46 well plate.

4.3.8 Cell Washes

During the course of collecting the samples for transcriptomics all stages were collected: the supernatant from the adhesion assay, the PBS wash and the trypsin wash. At each stage the numbers of cells collected decreased dramatically (Figure 4.11). Between the initial number of cells added to the flask and the number recovered in the trypsin wash fraction there is a 100-fold decrease in cell number, from over 625 $x10^4$ initially to 6-7$x10^4$. In order to obtain reliable RNA-Seq data a larger number of cells would be required, this led to the move from 6-well plates to T25 cell culture flasks in order to maximise the number of cells in the trypsin wash. The trypsin was stage was deemed most important for the transcriptomic experiments as this fraction included the trichomonad cells that had been physically attached to the monolayer and so would likely be expressing host-parasite attachment specific genes.

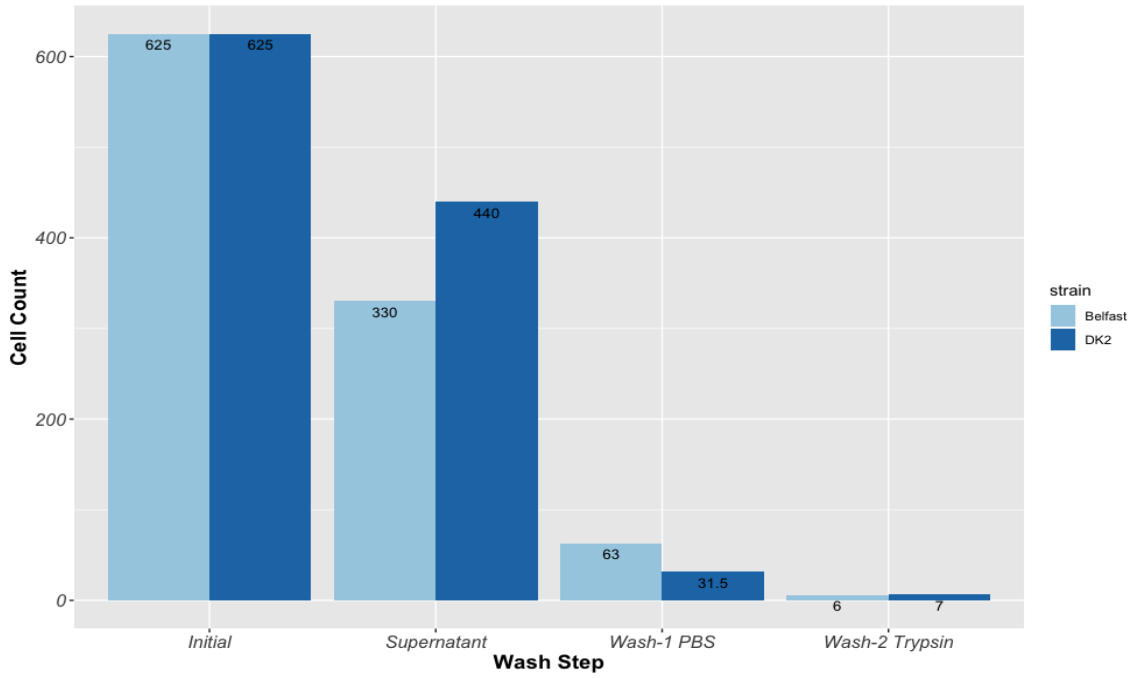

Figure 4.11: Numbers of *T. foetus* Belfast and DK2 cells initally added to the MDCK monolayer in a 6-well plate and per each cell wash fraction: supernatant removed from the well, 1st wash (PBS) and 2nd wash (trypsin)

4.3.9 RNA Isolation and Preparation

Due to the lack of integrity in some of the RNA samples, leading to them being processed using a different library preparation kit, there appeared to be a very large batch effect with the different preparation methods driving 94% of the variance. Corrections were made in DESeq2 for batch effect but the difference in variance was too high to remove the effect completely.

When all *T. foetus* DK2 samples were compared to identify the effect, if any, that the change in library preparation had on the results a clear batch effect was seen between those produced using the 'Zymo' kit rather than by poly-A selection (Figure 4.12). There was a variance of 94% when the first principle component is used. Due to this large batch effect the 'Zymo' samples were removed from the future analyses so only the poly-A selected samples were used. It was decided that the 'Zymo' samples should be removed as these came from the RNA samples with the lowest integrity and quality. Consequently, this low quality led to low numbers of reads being produced and low mapping percentages to the *T. foetus* genome (Table C1).

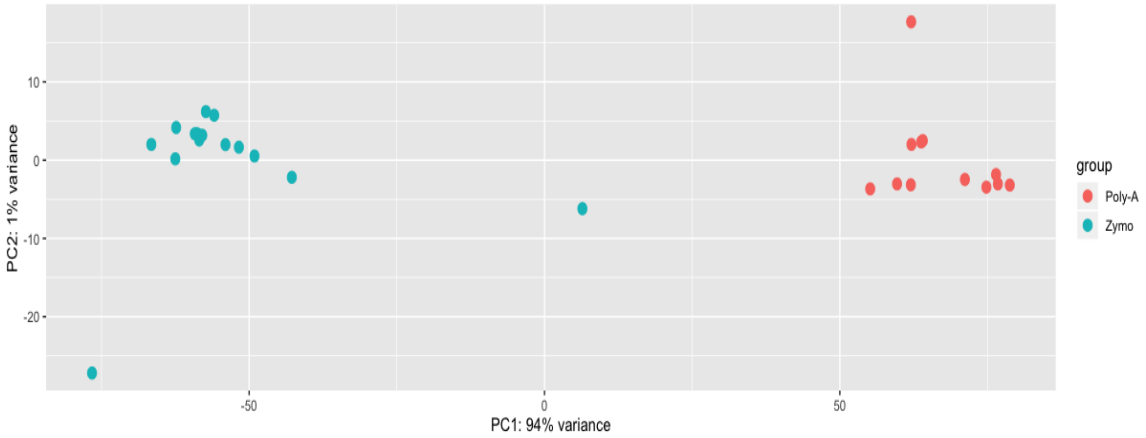

Figure 4.12: PCA plot produced in R with the DeSeq2 using principle components 1 and 2 showing the variance between samples. All samples are from *T. foetus* strain DK2 monolayer experiments where the DK2 cells were placed on an MDCK cell monolayer for varying lengths of time and the different washes were obtained. RNA was extracted from the samples and was processed either using poly-A selection or using a 'Zymo' kit. The samples treated using poly-A selection are red and those treated using the the 'Zymo' kit are blue.

There were also differences between the number of reads obtained for each sample and the mapping percentage to the produced and annotated *T. foetus* genome (Table C1). The number of reads ranged from 4,146,360 for the trophozoite trypsin wash to 51,657,211 for a DK2 only control, over a ten fold change difference.

4.3.10 Alignment Rates-Mapping Percentages

The mapping of the RNA-Seq reads to the *T. foetus* genome also varied hugely, from 0.04% to 95.9% (Table C1). The low value of mapping from the MDCK sample is to be expected, as MDCK cells are completely different to *T. foetus* cells. However, the mapping of the other samples, the controls and the washes, were expected to have a much higher mapping percentage in general.
The low percentages in some samples and not others, for example the range of 8.03% to 95.90% amongst the 1 hour trophozoite trypsin washes, suggests that either too few *T. foetus* cells were obtained or that the samples were contaminated with the MDCK host cells, this seems consistent, particularly amongst many of the trypsin washes. Additionally, it also seems that those samples with the lowest alignment rates are the 'Zymo' prepared samples rather than the poly-A selected samples. These samples had a lower RNA integrity before library preparation which may also explain the lower alignment rates.

4.3.11 Zymo Samples Removed

Due to the clear batch effect between library preparation methods, it was decided that the 'Zymo' produced samples were to be removed and analyses performed only on the poly-A selection samples. The poly-A selection samples were chosen as all other RNA-Seq analyses had been performed on samples with this library preparation method. Furthermore, the 'Zymo' produced samples came from lower quality RNA and showed low alignment rates to the genome and so very few gene were likely to be identified. When all 'Zymo' produced samples were removed, the analyses were rerun and the samples showed clear differences between them, although the variance of principle component 1 had been reduced to 50%. The samples cluster according to wash step, with the PBS samples appearing to be the most far removed from the other samples (Figure 4.13). The trypsin samples cluster together according to time points: 30 minutes and 1 hour.

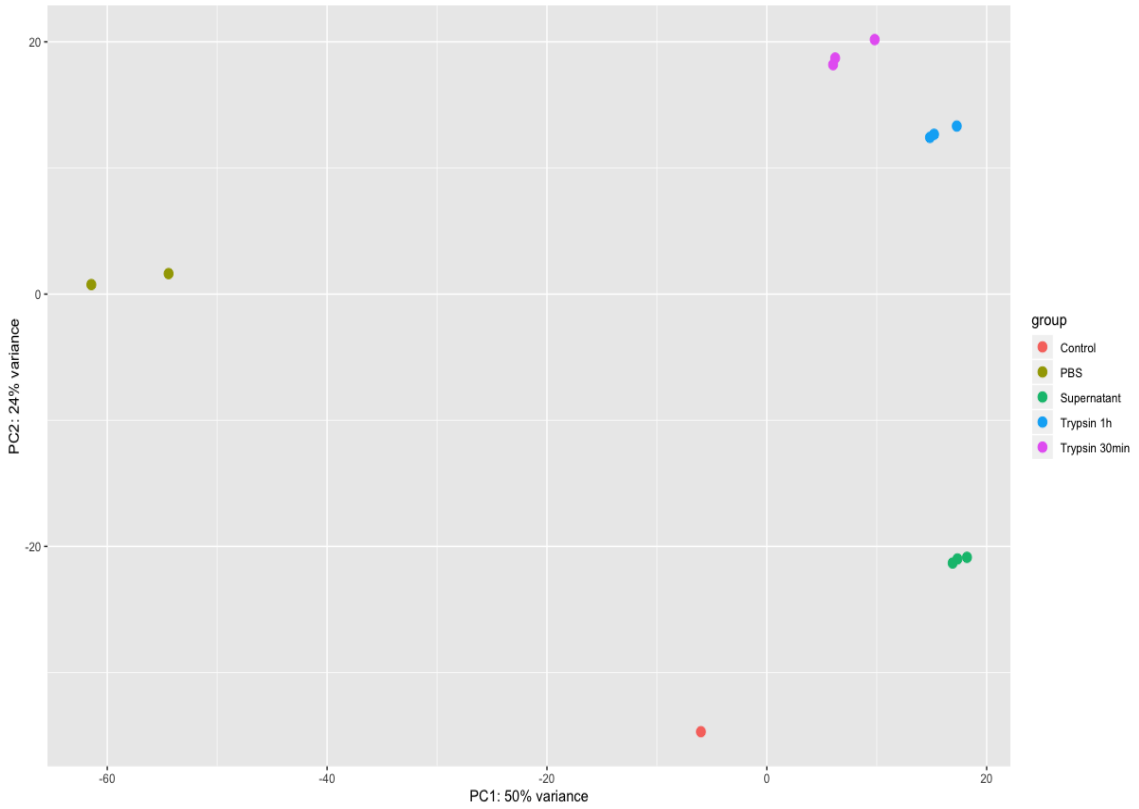

Figure 4.13: PCA plot produced in R with the DeSeq2 using principle components 1 and 2 showing the variance between samples. All samples are from *T. foetus* strain DK2 monolayer experiments where the DK2 cells were placed on an MDCK cell monolayer for varying lengths of time and the different washes were obtained. RNA was extracted from the samples and was processed either usin poly-A selection or using a 'Zymo' kit. All samples processed using the 'Zymo' kit were removed and the remaining samples were compared. The control sample is in red, trypsin wash samples are pink (30 minute samples) and blue (1 hour samples), PBS are yellow and supernatant are green.

After removing the 'Zymo' samples it becomes much easier to see the differences between the different wash stages with clear similarities in the expression profiles, for example, the PBS washes look very similar to one another, with the same genes more highly expressed, but different to the supernatant samples. The Trypsin 1 hour samples produce one cluster, as do the PBS, supernatant and 30 minute trypsin samples. However, removing all of the 'Zymo' samples led to a lack of replication in some instances, for example, there is now only one DK2 control. The trypsin samples for pseuodcyst and trophozoites also had to be combined so that there were three samples to compare. As shown in Figure 4.13, all the trypsin samples are similar regardless of which life stage they were originally from. Furthermore, as the samples were all incubated at 37°C it is highly likely that the low temperature induced pseudocysts would have reverted back to their trophozoite form quickly,sometimes within 10 minutes, therefore, all samples would essentially have come from trophozoites.

4.3.12 Comparison of DK2 and Belfast Negative Control Samples

Due to there only being one DK2 control that was processed using poly-A selection, we considered using the previously produced Belfast controls instead. The five Belfast controls and the single DK2 control were compared using DeSeq2 to identify if there were significant strain differences.

There appear to be large differences in variance (77%) using principle component 1 (Figure 4.14) between the Belfast and DK2 controls, whereas between Belfast samples themselves there is a much lower variance of 19%.

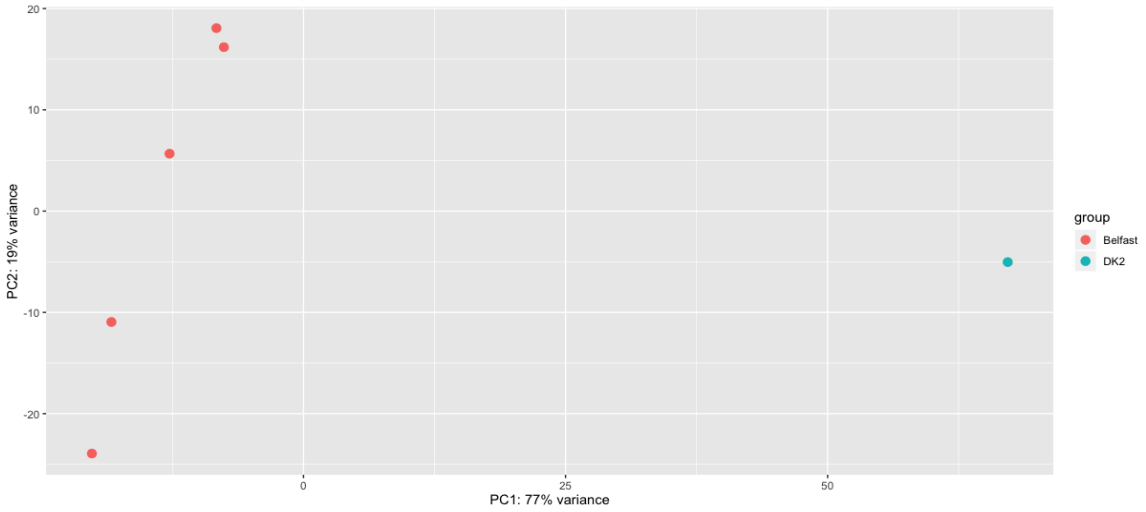

Figure 4.14: PCA plot produced in R with the DeSeq2 using principle components 1 and 2 showing the variance between samples. Samples are comparing control *T. foetus* trophozoites of strains Belfast and DK2. Cells were grown in Diamond media to log growth phase and then RNA was extracted.

Furthermore, when looking at the table of significantly differentially expressed genes (Table 4.4) from the RNA-Seq results there are 2,376 differentially expressed genes, with a fold change above 2 and a -log10 p value >0.01. There are 1221 significantly upregulated genes and 1155 significantly downregulated genes. The fold changes range from -14 to 15. As there were many significant differences between the two strains, the Belfast and DK2 controls were used to identify differentially expressed gene separately.

Comparison	No. Differentially Expressed Genes	No. Upregulated Genes	No. Downregulated Genes	Fold Change Range
DK2 Control vs Belfast Control	2,376	1,221	1,155	-14 to15
Trypsin Samples vs DK2 Control	5,353	2,625	2,728	-13 to 14.5
PBS Samples vs DK2 Control	5,038	2,625	2,728	-14 to 15
Supernatant Samples vs DK2 Control	6,266	3,659	2,607	15.7 to 11

Table 4.4: Number of significantly differentially expressed genes between transcriptome comparisons. The number of differentially expressed genes and whether these genes were up or down regulated were identified using DESeq2 in R. Genes were classed as significantly differentially expressed if -log10 pvalue>0.01 and fold change greater than 2. The range of fold changes from the most upregulated and downregulated genes per comparison are also shown.

4.3.13 Comparison of Trypsin Samples to DK2 Control

The DK2 controls were compared to the DK2 trypsin samples. These were the cells that had been attached to the MDCK monolayer and had not been removed when the supernatant was aspirated or when the monolayer was washed with PBS.

The results of the PCA plot using the DK2 control shows a large variance between the control and the trypsin samples, with a smaller variance between the trypsin time-point samples themselves (Figure4.13).

Regarding the number of significantly differentially expressed genes (Table 4.4) there is a larger range of fold changes, from -13 to 14.5. There are 5,353 significantly differentially expressed genes with 2,625 upregulated and 2,728 downregulated.

There are differences in the upregulated genes with CAMK protein kinases being upregulated and Iron-only hydrogenase subunits (Table C2). There are also no adenylate and guanylate cyclase catalytic domain containing proteins being upregulated, this suggests that this is a difference between the *T. foetus* strains rather than due to the experimental differences. The downregulated proteins have some similarities to those from the Belfast control samples with DnaJ, cysteine proteases and sialidase genes all being downregulated. However, there are differences: again there are no adenylate and guanylate cyclase catalytic domain containing proteins, CLN3 proteins or hydroxyglutaryl-CoA dehydratases.

4.3.14 Other Sample Comparisons to DK2 Control

Comparison of PBS wash sample to DK2 control

The PBS samples show an even greater variance between themselves and the DK2 control than the trypsin samples. They also show a large variance compared to the trypsin and supernatant samples (4.13).

In total there were 5,038 differentially expressed genes with 2,961 upregulated and 2,077 downregulated (Table 4.4). The fold changes are again, large, reaching -14 to 15 (Figure 4.4). CAMK proteins and hypotheticals were again, highly preferentially expressed in the PBS samples, with fold changes over 12 from the control. Iron-only hydrogenase was upregulated as was dihydropyrimidase (Table C3).

Comparison of Supernatant Samples to controls

The DK2 control and supernatant sample have a smaller variance than the other wash samples. There is also a small variance between the different replicates in the sample itself (Figure 4.13).

There are 6,266 differentially expressed genes, 3,659 upregulated in the DK2 supernatant samples and 2,607 downregulated (Table 4.4). The maximum fold change for upregulated genes is larger than the downregulated genes, 15.7 compared to 11.

The upregulated genes include CAMK kinases and serine proteases, similar to the other DK2 control comparisons. 5 of the top 10 downregulated are hypotheticals, similar to the DK2 trypsin wash (Table C4).

4.3.15 Comparison of Control and Experimental Samples in DK2

Between the two washes and supernatant samples, there were several gene products that were preferentially expressed relative to the control DK2 sample. Along with hypothetical genes, tubulin was preferentially expressed in each experimental condition relative to the DK2 controls, showing that cell cytoskelton proteins are upregulated in the presence of host cells. Protein binding proteins and transmembrane transporters were also upregulated in all experimental conditions as were catalytic activity related proteins. CMGC family protein kinases and CAMK genes are also upregulated

in all environmental conditions. CAMK genes are known virulence factors in *T. vaginalis* so it is likely they play a similar role in *T. foetus*.

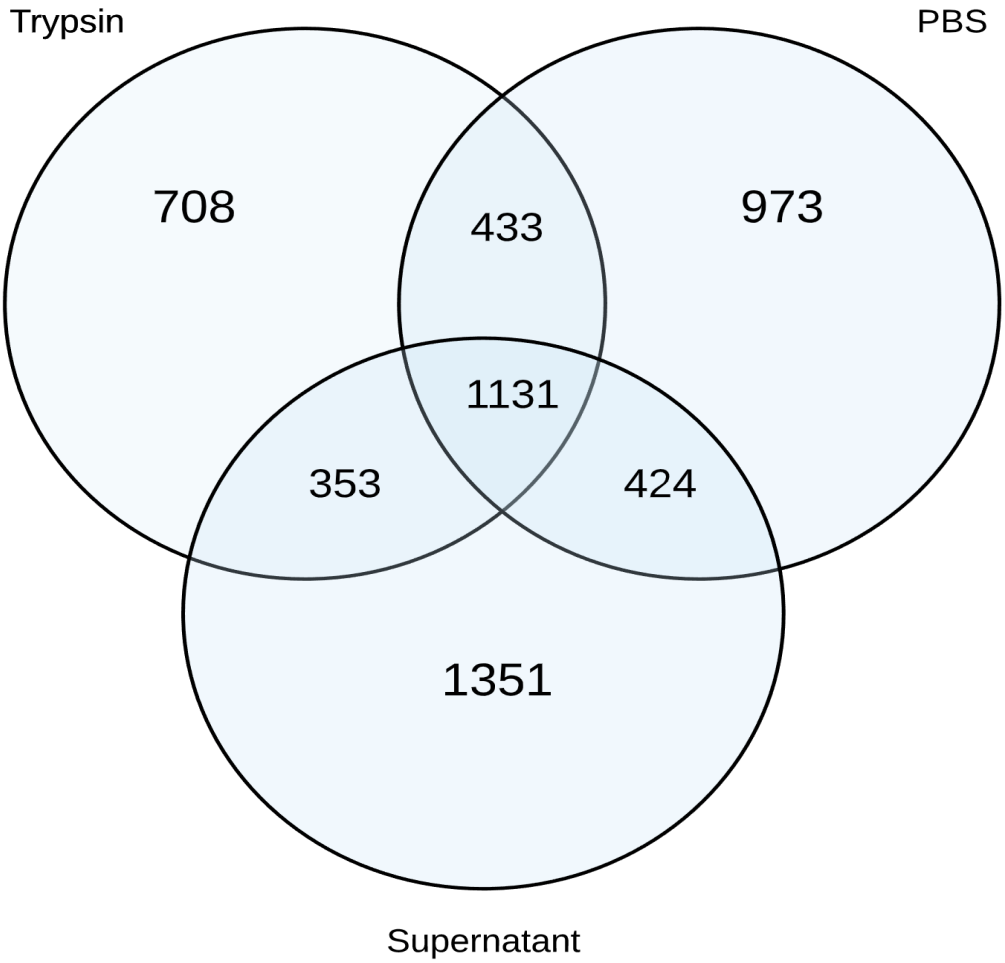

Figure 4.15: Venn diagram showing the number of preferentially expressed genes in the DK2 trypsin, PBS washes and supernatant samples relative to the DK2 control. For the wash and supernatant samples, DK2 cells were co-cultured with an MDCK monolayer and the parasites were then removed in the different wash steps. RNA was extracted and sequenced and differential expressed was identified using DeSEq2.

Over 1000 genes were preferentially expressed across all samples relative to the DK2 controls (Figure 4.15) with 353-433 genes preferentially expressed in two of the experimental sample conditions. Overall there were more genes that featured in two or more conditions than only one. The supernatant samples had the most genes that were specifically upregulated compared to the other samples, potentially because these were not cells that had necessarily been in contact with the monolayer and so had not expressed some host specific genes.

Immunodominant variable surface antigen like proteins were found upregulated in the trypsin and PBS samples (7 genes and 9 genes respectively) rather than the supernatant samples. As no immunodominant variable surface antigen like proteins were found in the supernatant or control this suggests that they require actual attachment to a host cell to increase expression. Adenylate and guanylate cyclase catalytic domain containing protein was also downregulated in all samples, this is involved in the conversion of GTP to GMP [427] and so a reduction in the expression of these proteins may also lead to a reduction in metabolic activity, potentially as the parasites are obtaining nutrients from the host cell monolayer.

4.3.16 Comparison of Differentially Expressed Genes and Peptide Array Proteins

When the 50 proteins used in the peptide array in Chapter 3 were compared with the differentially expressed DK2 genes from each wash, 19 proteins were found differentially expressed in at least one of the samples. Eleven proteins were preferentially expressed in the experimental samples relative to the DK2 controls and 7 were preferentially expressed in the DK2 controls.

Three proteins: TTF57869, TTF09721 and TTF36584 were found in the upregulated trypsin wash samples suggesting that they are upregulated by coming into contact with host cells. TTF09721 was also found to have the top 20 highest number of immunogenic epitopes for the natural infected samples. Five proteins (TTF14901, TTF38179, TTF347783, TTF16447 and TTF45090) were found in both washes and the supernatant, suggesting that the presence of host cells increases their expression. Of those five, three (TTF14901, TTF38179 and TTF34783) were found in the top 20 proteins for the number of expressed epitopes. Two (TTF11874 and TTF75903) were found in both the PBS and trypsin samples, again suggesting that some form of contact with the host cells leads to an increase in expression.

Of those peptide array proteins that are downregulated, four are found in the supernatant only (TTF36321, TTF35905, TTF12012 and TTF41308), two were found in all washes and supernatant (TTF14449 and TTF44685) and one was found in the trypsin wash only (TTF11197) meaning that the presence of host cells reduces their expression. This would mean that they are less likely vaccine candidates as they are not naturally expressed in the presence of the host.

4.4 Discussion

In this chapter cell assays were performed on five strains of *T. foetus*: Belfast, DK2, BP4, KV1 and UT and one strain of *T. mobilensis* to discover whether different strains had different cytotoxic and adhesion properties. One strain, DK2, was used in subsequent experiments to obtain RNA from cells that had been grown in the prescence of an MDCK monolayer. *T. foetus* cells from the supernatant, PBS wash and trypsin wash from the co-culture had RNA extracted and sequenced and preferentially expressed transcripts between the stages and between the co-culture and DK2 cells only were obtained. The preferentially expressed genes from *T. foetus*/MDCK co-cultures were cross-checked with the peptide array results from Chapter 3 to create a list of genes which could be tested for structural invariance in Chapter 5.

4.4.1 Limitations of Cell assays

To identify dead cells during the comparison of *T. foetus* strains experiments, trypan blue exclusion assays were used. However, the samples were only taken from supernatant so dying cells in the monolayer were not counted and pieces of cell debris could have been from the *T. foetus* cells rather than the MDCKs. To account for this only whole dead cells were counted rather than debris. MDCK cell detachment was a key measure of monolayer destruction and so the supernatant would be the best sample to measure of this.

One other limitation is that trypsin-EDTA was required to detach the *T. foetus* cells from the MDCKs. This could potentially mean that the differential gene expression could be due to the presence of trypsin rather than the presence of the host cell. However, the fact that the same genes also appear to be preferentially expressed in the supernatant and PBS cell fractions, suggests it is caused by the presence of the host cells rather than the presence of the trypsin-EDTA.

4.4.2 Monolayer Cell Death

There has been debate on whether trichomonad infection causes damage to the host cells by attaching to them [420] [421] or whether it is due to something that the parasites are secreting, such as proteases [414]. The results show (Figures 4.4, 4.7 and 4.8) that the presence of the trichomonads causes damage.

There seem to be three potential reasons for the increase in death rate of the MDCKs when *T.*

foetus cells are added to them.

1) The *T. foetus* cells are producing secretions that are cytotoxic to the MDCKs
2) The *T. foetus* cells are using up all the nutrients in the media, causing the MDCKs to be nutrient deficient leading to cell death.
3) The *T. foetus* cells are producing waste products which are toxic to the MDCKs.

Due to the fact that the increase in cell death of MDCKs occurs within 2 hours and the rate of cell death is different between strains, it is more likely the *T. foetus* cells are producing enzymes and other secretions that are cytotoxic. If the cell death was due to a lack of nutrients or build-up of waste products, the cells death would increase at a later time point, such as 24 hours, as is the case in the control samples. Furthermore, the death rate would be consistent between strains as the same number of starting cells were used.

The damage to the MDCK monolayers when the Belfast strain *T. foetus* cells were added was lower compared to all the other strains. This may be due to the fact that it has been grown in culture for many years and passaged numerous times, and the results are similar to those reported by De Jesus *et al.* (2009) [170] who identified that fresh clinical isolates of *T. vaginalis* were more cytotoxic and damaging to cell monolayers than those grown in culture for long periods. The fresh isolates used by De Jesus *et al.* also had higher enzyme activity than the cultured strains [170].

Curvo *et al.* (2008) [171] compared strains of *T. vaginalis* with differences in virulence and looked for proteomic differences between them. Of the genes that were found to have increased expression in the more virulent phenotype several were also found upregulated in the more destructive DK2 strain of *T. foetus* compared to the Belfast strain in my transcriptome experiments. Seven Clan CA family proteins were found to have increased expression in the DK2 strain. Clan CA proteins were also found to have increased expression in the more virulent *T. vaginalis* strain [171]. Haloacid dehalogenase-like hydrolase family protein and cathepsin like cysteine protease were also found to be preferentially expressed in both the virulent *T. vaginalis* strain and DK2 [171] suggesting they are involved with increases in virulence or cell damage.

De Miguel *et al.* (2010) [161] also compared fresh *T. vaginalis* strains to those grown in long-term lab culture to identify potential virulence associated proteins. Of the proteins they identified as preferentially expressed in the more virulent strain, two categories were also found preferentially expressed in the DK2 strain: serine-threonine protein phosphatase of which 17 genes were upregulated and adenylate and guanylate cyclase catalytic domain containing protein, of which 16 were

upregulated in the DK2 control. Of the proteins identified in the less virulent *T. vaginalis* strain [161], two categories were also found preferentially expressed in the less cytotoxic Belfast controls: ABC transporters of which 13 genes were upregulated and auxin efflux carrier proteins, of which 4 were upregulated in the Belfast samples.

4.4.3 Differentially Expressed Genes in the Presence of a Host Cell Monolayer

When the differentially expressed genes from the controls were compared to the different wash samples, there were more similarities between the washes than between the washes and controls. In all experimental wash samples for example, cytoskeleton related genes and CAMK proteins-known virulence factors, were upregulated. This was also the case when the Belfast controls were used rather tha th DK2 control, suggesting that these genes are not strain specific and are upregulated in the presence of the host monolayer. There are gene that are upregulated in all of the experimental samples but not the controls, such as ankyrin repeats, adhesin-like proteins and CAMK proteins, suggesting that these are preferentially expressed in the presence of host cells, even if the parasites do not have direct contact with the monolayer. Key gene families that were upregulated included: CAMK and cysteine proteases, Ser-Thr kinases and leucine-rich repeat containing proteins.

However, due to the issue with producing the samples, there is only one control replicate with which to compare the different environmental samples so results may not be fully representative. The results do, however, seem in agreement with past literature as several of these genes have been associated with virulence and pathogenicity in trichomonads [67] [418], particularly the CAMK genes.

CAMK and Cysteine Proteases

CAMK and cysteine proteases are known virulence factors in a variety of organisms [428], having roles in autophagy, adhesion and penetration of host tissues. In *Entamoeba histolytica* they are known to degrade the extracellular matrix and it has been postulated that they can degrade IgA [428] in addition to invasion and parasite encystation. In trypanosomes, they are known to be virlence factors and have roles in evading the host immune system and iron aquisition [428].

Ser/Thr Kinases

Ser/Thr protein kinases are signalling proteins in eukaryotes [429] and can modulate signal transduction pathways. They were found to be preferentially expressed in *T. foetus* in the presence of a host cell monolayer and so may have roles in virulence or pathogenicity as, in other organisms, they possess these functions. They are also known virulence factors in bacteria [430] [429] where they have roles in environmental sensing and evading the host defences.

In *Yesinia* species Ser-Thr kinases have been found to affect the host cytoskeleton, particularly disrupting actin [429] and in *Shigella* they have been found to downregulate the host immune system. In *Streptococcus pyogenes* Ser/Thr protein kinases have been found to be required to induce disease in a mouse model [430] and can activate virulence genes. Additionally, they were found to be involved in regulation of metabolism, cell division and penicillin tolerance [430]. In *Toxoplasma* they have been implicated in host invasion [431]. It is therefore possible that similar functions could be performed by trichomonad Ser-Thr kinases, explaining why they are preferentially expressed in the presence of the host cells.

Leucine-rich Repeats

As stated in previous chapters, the BspA family are leucine-rich repeat containing proteins. Leucine-rich repeats are involved in recognition of cell surface motifs [166]. BsPA proteins are known to be involved in adherance to host cells [167] and pathogenicity [165] [33] and so the increase in preferential expression in the presence of the host are in keeping with this.

4.4.4 Comparison of Differentially Expressed Genes to Cell Surface Predictions

Of those proteins that feature in the network, CAMK and leucine rich repeat containing proteins also appear in the preferentially expressed genes when the parasites are in contact with a host monolayer as do immunodominant variable surface antigens. All three groups have been associated with trichomond-host interactions [67] [68] particularly the cysteine proteases. These have been suggested as vaccine candidates previously for *T. vaginalis* [68]. Additionally the immunogenic proteins from the peptide array have been found, in some instances, to also be preferentially expressed in the presence of the host cells, such as TTF14901 and TTF38179. There were also a large

number of hypotheticals preferentially expressed across the different washes: 971 across the trypsin washes, 1,203 across the PBS washes and 1,296 across the supernatant samples. As these appear to only be present in *T. foetus*, these again would make interesting potential vaccine candidates if they follow the other criteria.

To create the list of candidate genes for Chapter 5, the top 20 proteins, with the highest numbers of epitopes and maximum intensity, that were found in the peptide array of Chapter 3 were selected. These proteins had high number of epitopes and maximum intensities in the presence of both the experimentally infected and naturally infected sera. Additionally, the genes that were found to be upregulated in all of the trypsin, PBS and supernatant samples were examined. These were then cross-checked with the data from Chapter 1 to identify which of these genes also possesses one transmembrane helix. This produced the final list of 77 genes that are going to be examined for SNPs and differences within them between the different *T. foetus* species (Table 4.5).

Protein ID	Product
TTF00910	hypothetical protein TRFO_06137
TTF01152	hypothetical protein TRFO_30327
TTF01161	hypothetical protein TRFO_36810
TTF02039	hypothetical protein TRFO_30849
TTF02745	hypothetical protein TRFO_08363
TTF03161	hypothetical protein TRFO_16017
TTF03169	hypothetical protein TRFO_16005
TTF04635	hypothetical protein TRFO_32884
TTF05043	—NA—
TTF05727	hypothetical protein TRFO_04170
TTF05864	hypothetical protein TRFO_16898
TTF08006	hypothetical protein TRFO_19408
TTF08916	hypothetical protein TRFO_04830
TTF09063	hypothetical protein TRFO_39123
TTF09721	adhesin-like protein
TTF11197	hypothetical protein TRFO_10561
TTF12012	hypothetical protein TRFO_39497
TTF13670	hypothetical protein TRFO_19601
TTF13877	hypothetical protein TRFO_19801
TTF14449	hypothetical protein TRFO_11445

TTF14901	hypothetical protein TRFO_36551
TTF16365	cation channel sperm-associated protein subunit beta-like
TTF16447	hypothetical protein TRFO_30251
TTF18002	hypothetical protein TRFO_42877
TTF19171	PKD domain-containing protein
TTF20128	LPXTG-motif cell wall anchor domain protein
TTF21569	DUF2804 domain-containing protein
TTF21670	ankyrin repeat protein
TTF23071	hypothetical protein TRFO_15320
TTF25642	T9SS C-terminal target domain-containing protein
TTF28036	hypothetical protein TRFO_12081
TTF28414	hypothetical protein TRFO_36741
TTF29007	hypothetical protein TRFO_03744
TTF31219	hypothetical protein TRFO_20571
TTF33095	hypothetical protein TRFO_42895
TTF33122	Mannosyl-oligosaccharide 1,2-alpha-mannosidase MNS2
TTF33823	hypothetical protein TRFO_22464
TTF33905	hypothetical protein TRFO_09572
TTF34783	hypothetical protein TRFO_09395
TTF37306	hypothetical protein TRFO_41775
TTF38179	hypothetical protein TRFO_27265
TTF40724	hypothetical protein TRFO_38622
TTF43413	two component sensor and regulator histidine kinase bacteria
TTF43627	hypothetical protein TRFO_32949
TTF44528	adhesin-like protein
TTF44600	conserved domain protein
TTF44685	hypothetical protein TRFO_33209
TTF44949	hypothetical protein TRFO_32080
TTF45042	adhesin-like protein
TTF45771	hypothetical protein TRFO_26599
TTF46221	tyrosine kinase, putative
TTF47476	hypothetical protein TRFO_08795
TTF48521	hypothetical protein TRFO_11786
TTF49933	hypothetical protein TRFO_12226

TTF50636	Lecithin:cholesterol acyltransferase family protein
TTF53402	hypothetical protein TRFO_12669
TTF56587	hypothetical protein TRFO_32199
TTF62174	hypothetical protein TRFO_11993
TTF62376	hypothetical protein TRFO_43242
TTF63395	hypothetical protein TRFO_33646
TTF63680	adhesin-like protein
TTF67401	hypothetical protein TRFO_35350
TTF70906	hypothetical protein TRFO_12915
TTF70954	hypothetical protein TRFO_11878
TTF71197	putative outer membrane protein pmp20 precursor
TTF72511	hypothetical protein TRFO_29759
TTF72966	hypothetical protein TRFO_23368
TTF73221	hypothetical protein TRFO_39563
TTF73824	hypothetical protein TRFO_11378
TTF74586	hypothetical protein TRFO_16588
TTF74797	hypothetical protein TRFO_03869
TTF75601	adhesin-like protein
TTF79943	hypothetical protein TRFO_14199
TTF80312	hypothetical protein TRFO_42647
TTF82375	Synaptobrevin family protein
TTF83271	hypothetical protein TRFO_31294

Table 4.5: Shortlist of genes to be tested for structural invariance in Chapter 5. These proteins either feature in the Chapter 3 peptide array where they showed high maximum intensity and large numbers of significantly immunogenic epitopes; or they were found preferentially expressed in the co-culture transcriptome samples and possessed a transmembrane domain

4.4.5 Future Work

Future work could involve looking at the transcriptome of the MDCK cells, both axenic and after *T. foetus* strains are added, to identify which genes are preferentially expressed due to the host-parasite interactions. This could provide a more targeted approach to identifying potential vaccine candidates. The other *T. foetus* strains could be examined and gene expression could be compared to the Belfast strain to identify other strain specific preferentially expressed genes. This may also allow further comparisons between the expression of certain genes and the virulence and cytotoxicity

of the strain, in a similar way to *T. vaginalis*. The different strains could also be tested for differentially expressed genes in the presence of a host cell monolayer to identify whether the genes seen in the DK2 and Belfast strains are representative of all strains. This could also be tested using *T. mobilensis* and *T. vaginalis* to identify if genes identified are related to virulence in several trichomonad species or just *T. foetus*. If the same genes were found to be related to cell adhesion and cytotoxicity in all of the trichomonad strains tested then there could be a potential for a more generic trichomonad vaccine.

4.4.6 Conclusion

In this chapter, five different strains of *T. foetus* along with *T. mobilensis* were co-cultured with a monolayer of MDCK cells and differences in their adherence and cytotoxic properties were observed. The transcriptomes of the DK2 strain, which appeared to be the most destructive, were produced for several washes in the presence of the MDCK monolayer. These transcriptomes from *T. foetus* cells, following co-culture were compared to both a DK2 control and Belfast controls and genes that were preferentially expressed in the presence of the monolayer, identified. These genes were compared to the peptide array proteins in Chapter 3 and Network proteins in Chapter 2 to identify key gene families that appeared to be associated with virulence and pathogenicity. A subset of 77 genes has been found which fulfils all of these criteria. The next stage is to identify which genes in *T. foetus* are structurally invariant between the populations, whcih is the aim of chapter 6, and then to cross check these with the peptide array, cysteine peptidases and variable surface antigens to produce a short list of potential vaccine candidates (Figure 4.16).

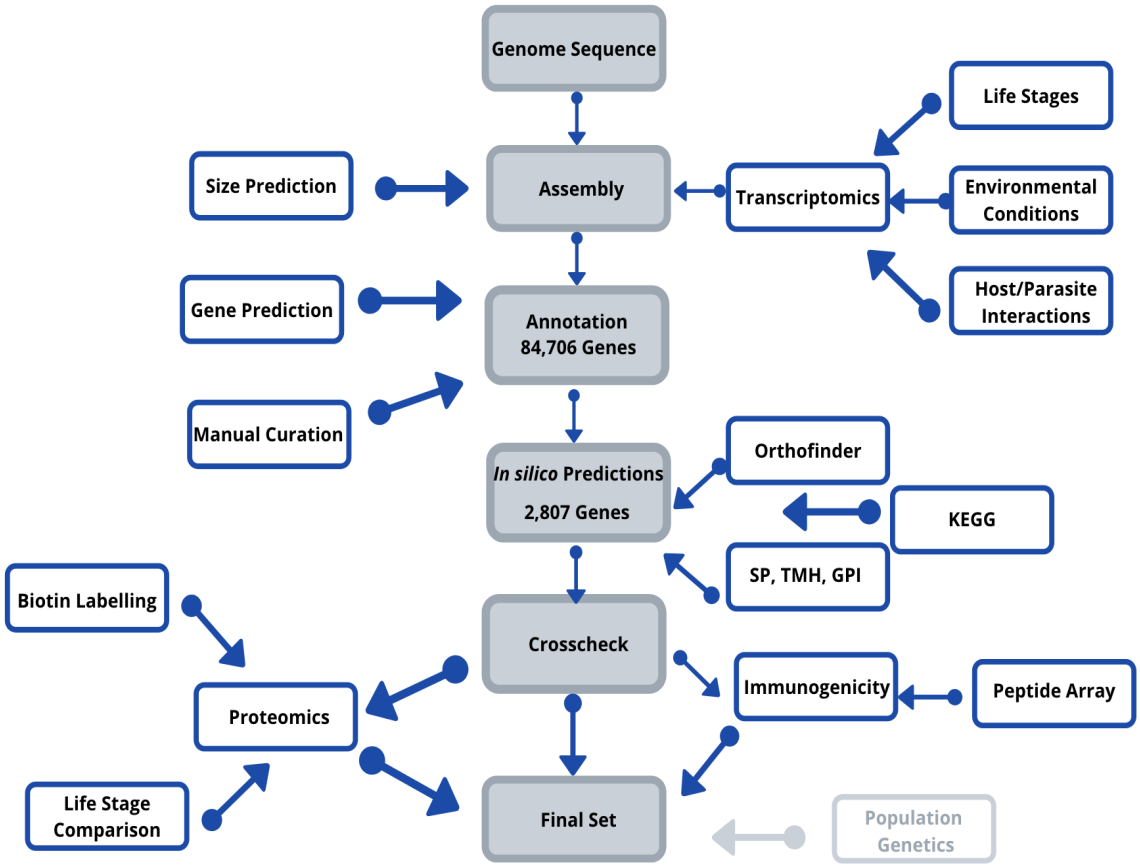

Figure 4.16: Reverse Vaccinology Project Flow Diagram. The sections highlighted are those that have been completed over the course of this chapter and previous chapters and those that are pale are to be completed in subsequent chapters.

Chapter 5

Population Genetics of *T. foetus* and *T. mobilensis*

5.1 Introduction

When creating a vaccine and identifying vaccine candidates, one of the key requirements is that it must protect against all strains of the pathogen.

In some parasites, such as *Theileria* [432], there are strain differences that have lead to issues in vaccine design and production. The vaccine produced against this parasite uses a cocktail from three isolates in order to cover more genetic diversity [432]. In a similar way, a *T. foetus* or trichomonad vaccine may need multiple genetic variants from the strains in one cocktail to produce the most effective vaccine. However, the *T. foetus* strains used in my analysis are not from Africa or Asia so there is relatively little geographic genetic diversity, the strains are from North America, Ireland and Czechoslovakia. The *Theileria* vaccine itself is also non-standardised [433] meaning that between different batches of the vaccine there can be slightly diffrent components. In *Theileria*, p67 [434] was used to inoculate cattle. However, only 70% of them were immune when later challenged with a buffalo strain of *Theileria*. The buffalo strain is 95% identical to the cattle strain, however, it does contain a peptide insert.

Amzati *et al.* (2019)[435] identified multiple polymorphisms in the *Theileria* genes Tp1 and Tp2, genes that were potential vaccine candidates. Six variants were found in Tp1 and 9 in Tp2 caused by 10 and 11 different alleles respectively [435]. The diversity appeared to depend on the geographic location of the *Theileria* isolates tested, i.e. isolated from lowland or highland areas. A trivalent

vaccine is used for *Theileria* to attempt to combat this, commonly seen, geographic diversity. If diversity is seen between *T. foetus* strains from different continents in the candidate genes, a bivalent or trivalent vaccine may be needed to account for the polymorphisms and diversity.

A bovine trichomoniasis vaccine needs to work against all *T. foetus* strains circulating in nature; if the vaccine candidate is significantly different between strains then the vaccine might be unsuccessful in certain populations or multiple vaccines would need to be created. Furthermore, if the vaccine candidate is highly polymorphic, it is possible that it could evolve or mutate further, potentially making the vaccine less effective over time.

Thus, we must select a candidate antigen that confers immunity against all strains of *T. foetus*. In this chapter, I employ genome re-sequencing of *T. foetus* strains to quantify single nucleotide polymorphisms (SNPs) across the 'short-listed' loci to identify those genes that are most conserved among different *T. foetus* strains and trichomonad species.

5.1.1 SNPs and Indels

There are many variables that can alter the sequence of a genome: large scale variants such as transposons, tandem repeats and duplication events and small scale variants such as SNPs and indels.

Single nucleotide polymorphisms (SNPs) are a change to a single nucleotide base in a genetic sequence [436]. They comprise 80% of known genetic polymorphisms [437] and can affect the likelihood of acquiring certain diseases, along with affecting the likelihood of recovery [437]. They are also commonly used for genetic tests as they can serve as biomarkers [437]. Additionally, it has been found that SNPs that occur in introns and those that appear to be silent, i.e, changing a nucleotide which does not change the amino acid, can also affect the function of genes [437]. Insertion deletion variants (indels) add or remove one or more nucleotides from a sequence [436]. They can have impacts in coding and non-coding regions (Figure 5.1).

SNPs can also be used for genotyping and diagnostics [438] for instance, single locus genotyping has been used to identify different species of *Plasmodium* in samples, even when they show multiplicity [439]. The SNP panels produced can help in identifying which *Plasmodium* species are in the sample, even when low amounts of DNA are present. 24 SNP markers have been used to identify *Plasmodium falciparum* from patients and laboratory samples [440]. SNP biomarkers and single

locus genotyping have been used to identify speciation in *Rickettsia* in ticks [441]. This same procedure could be applied to different trichomonad species to determine if they are the same species, such as in the case of *T. foetus* and *T. suis* and could allow for more accurate phylogenetic reconstruction.

Figure 5.1: Example of SNPS and an indel in a genetic sequence. Sequence 2 contains two SNPs and one indel relative to reference sequence 1, highlighted in green and red respectively.

5.1.2 Synonymous and Non-synonymous Mutations

Amino acids show redundancy, meaning that different codons can code for the same amino acid, this leads to the rise of synonymous and non-synonymous mutations.

Synonymous mutations are mutations that do not change the amino acid coded for, for example, a change from CGA to CGG will encode arginine [442], therefore they are also known as silent mutations.

Non-synonymous mutations are mutations that do produce an amino acid change and are also known as substitution mutations. There are two types of non-synonymous mutations, missense and nonsense mutations. Missense mutations occur when one amino acid is changed to another, for example TGC to TGG would change the amino acid from cysteine to tryptophan. A nonsense mutation changes the amino acid to a stop codon, for example, changing TGC to TGA would change the cysteine to a stop codon.

Nucleotide Substitution-dN/dS ratios

Synonymous substitution has been shown to have very little effect on the organisms ability to survive [443]. The substitutions have been found to be more rapidly appearing in pseudogenes [443]. Synonymous differences are known as d_S and non-synonymous differences are d_N. d_N is usually smaller than d_S and a d_N/d_S ratio can be calculated. This ratio estimates how much non-synonymous evolution (non-neutral) has occurred compared to synonymous selection (neutral). If the d_N/d_S is 1, a gene or site is evolving neutrally. If the d_N/d_S ratio is <1, this suggests that non-synonymous mutations in a gene or site are selected against (purifying selection). Finally if the d_N/d_S ratio is >1, this indicates that the non-synonymous changes are being favoured (positive selection), and the gene is changing more quickly.

5.1.3 Population Genetics

Population genetics is the study of biological populations and the changes in genetic composition that occur within and between them [444]. It also aims to explain molecular evolution seen between populations in addition to their genetic variation [445].

As Tibayrenc *et al.* (1995) [446] mentions, there are well known issues in classifying different species within microorganisms. Population genetics assists in finding subdivisions within larger populations that may otherwise be missed. It can also be used to evaluate how genetic diversity affects diversity in a practical setting, e.g. drug resistance. *T. cruzi*, for example, was found to have two major lineages, each divided into subpopulations. The different subpopulations and their genetic distances had marked effects on several properties, such as parasitaemia in mice, growth rate in culture and infectivity [446].

Barry *et al.* (2009) [447] used 10 potential *Plasmodium falciparium* candidate antigens from three stages: pre-erythrocytic, merozoite and gametocyte. It was found that the merozoite antigens had high levels of diversity and this diversity was spread globally, whereas the diversity of the non-merozoite antigens was due to physical barriers and different geographic locations. There were high levels of diversity in countries with high levels of malaria transmission, such as those in sub-Saharan Africa, and lower levels in the Americas where transmission is known to be low.

The five strains of *T. foetus* will be sequenced using Illumina sequencing. Each strain originates either from a different location or a different time point of collection. Strain KV-1 originates from Czechoslovakia in 1962, strains BP-4 and UT from Beltsville, Maryland in 1956 and 1967

respectively. Strain DK-2 originates from Davis California in 1967 and the Belfast strain, that has been used as a reference, is from Belfast in 1938. As the strains of *T. foetus* are from different geographic locations, such as Maryland and Czechoslovakia, it would be interesting to see whether differences in genes is location specific. Furthermore, if a vaccine is created, it must be conserved in all strains across all continents. For example, the vaccine for *P. falciparum* mentioned by Barry *et al.*, [447] seems to be based on low prevalence haplotypes from one subgroup, rather than representing several global populations of the parasite. They also identified between 3 and 6 subgroups within the antigens. Haplotypes within the subgroups were similar and able to elicit antibody responses, whereas haplotypes between different subgroups were not similar and could not elicit the same responses. Furthermore, as the samples were collected decades apart in some instances, it would be interesting to identify if the year of collection affects the gene expression and mutations. Using samples from several locations and time points will maximise the potential variation seen so that vaccine candidates will be present across these strains. This would aim to create consistent responses by the vaccine candidates, regardless of where the original infection has been found. However, all strains used are laboratory strains that have been in lab culture for decades. The length of time the isolate has been in lab culture may also provide interesting results.

Huyse *et al.* (2005) [448] stated that the environment within the host changes quickly, which leads to more rapid speciation in parasites. There are also differences between each separate parasite population and there can be different speciation pressures within these subpopulations [448]. As the *T. foetus* strains are, again, from different areas, they may have had different selection pressures. This could be particularly marked if a comparison was made between feline and bovine strains. Additionally, parasites tend to have short generation times and fragmented populations [448]. Low host specificity and host switching can also affect phylogenetic timelines, for instance switching between cat and cow hosts in *T. foetus* or different bird species for *T. gallinae*. There is a large amount of diversity in natural infections which will affect the number of SNPs present between the strains and species. Comparing the five *T. foetus* strains and the *T. mobilensis* strain could give a good indication of how much natural variation there is within and between species that need to be taken into account when choosing the vaccine candidates.

Tajimas D

Tajimas D is a comparison for neutrality used in population genetics. It compares the average number of pairwise differences with the number of segregating sites [449] [450].

The equation Tajima's D is:

$$D = \frac{\hat{\theta}_{\mathrm{T}} - \hat{\theta}_{\mathrm{W}}}{\sqrt{\hat{V}(\hat{\theta}_{\mathrm{T}} - \hat{\theta}_{\mathrm{W}})}}$$

Where $\hat{\theta}_{\mathrm{T}}$ is the number of pairwise differences or π, V is the variance and $\hat{\theta}_{\mathrm{W}}$ is the number of segregating sites.

When Tajima's D is negative, positive selection is occurring, if the population is not undergoing changes such as migration or expansion. In this case there are more segregating sites than pairwise differences [449]. When the Tajima's D value is positive, balancing selection is occurring and there are more pairwise differences than segregating sites.

Bottlenecks can have large effects on the Tajima's D value, as the population expands and recovers the Tajima's D value drops sharply before eventually increasing [449]. This is due to the increase in population leading to a high rate in new mutations [451]. By looking at the Tajima's D value across the *T. foetus* genomes it may be possible to determine whether there has been a recent expansion of bottleneck in the populations history.

5.1.4 Variants in Trichomonads

As described in Chapter 4, different strains of *T. foetus* have different cytotoxic and adherence effects [414] [418], and in some cases these have been traced to individual genetic variants.

Paulish-Miller *et al.* (2014) [452] identified resistance in *T. vaginalis* to metronidazole has been associated with SNPs in two nitroreductase genes: $ntr4_{Tv}$ and $ntr6_{Tv}$. It was also suggested that these SNPs could be used as diagnostic tests to identify whether a *T. vaginalis* strain in an infection will be metronidazole resistant. It is possible that there are similar SNPs within *T. foetus* strains that confer the same resistance, possibly explaining the increase in antibiotic resistance seen in the species.

Bradic *et al.* (2017) [228] sequenced 100 isolates of *T. vaginalis* and identified 72 SNPs related to metronidazole resistance, predominantly in genes that were deemed 'hypothetical' and were generally enriched in genes showing gene ontology to drug detoxification. SNPs were also identified that showed convergent evolution between *T. vaginalis* and *T. foetus*. They also identified SNPs for genes that currently have no function.

5.1.5 Aims and Objectives

This chapter aims to identify the SNPs within the genomes of five strains of *T. foetus*. These will then be compared to the list of possible vaccine candidates from Chapter 4 to identify which of these genes are most conserved across all strains. There are four objectives.

1) The genomes of five *T. foetus* strains and one *T. mobilensis* strain will be sequenced using Illumina sequencing.
2) Map the newly created genomes to the reference PacBio Belfast genome
3) Call the SNPs found between each strain and the reference genome
4) Calculate the population parameters and statistics for each gene, such as Tajima's D and d_N/d_S ratio.

5.2 Methods

5.2.1 Trichomonad Strains Used

Multiple strains of *T. foetus*, along with related parasite, *T. mobilensis* were used. This was to identify how many SNPs are seen between each strain and the reference *T. foetus* genome and whether there are similarities between the different strains and species. In total, five *T. foetus* strains were used along with *T. mobilensis*:

1) *T. foetus* KV-1 strain-naturally infected from Czechoslovakia (1962) ATCC 30924
2) *T. foetus* BP-4 preputial washings from Beltsville Maryland (1956) ATCC 3003
3) *T. foetus* UT Preputial washings from Beltsville Maryland (1967) ATCC 30232
4) *T. foetus* DK-2 Preputial washings from Davis California (1967) ATCC 30231
5) *T. foetus* Belfast strain, Belfast, Ireland (1938) ATCC 30166
6) *T. mobilensis* USA-M776 Bolivian squirrel monkey (1984) ATCC 50116

In the same way as Chapter 4, cells were initially grown in ATCC entamoeba medium before being passaged into Diamond media. All cells were incubated at 35°C.

5.2.2 DNA Extraction

DNA was extracted from all six strains using the Qiagen 'All Prep DNA/RNA mini Kit' as per the manufacturers instructions. 100μl was of eluted DNA was produced The amount of DNA and the 260/280 and 260/230 ratios were all calculated using the nanodrop and concentration was calculated using Qubit.

Strain	Concentration (ng/μl)	260/280	260/230	Total DNA amount (μg in 100μl)
Belfast	191.8	2.11	0.71	19.18
KV-1	86.9	1.92	0.36	8.69
DK-2	166.9	1.97	0.93	16.69
BP-4	117.5	2.06	0.54	11.75
UT	155.5	1.90	0.34	15.55
Mobilensis	134.6	1.95	0.34	13.46

Table 5.1: DNA preparation statistics for one *T. mobilensis* and five *T. foetus* strains. DNA was extracted using the Qiagen 'All Prep DNA/RNA mini kit', the 260/280 and 230/280 ratios were calculated using the nanodrop and concentration was calculated using Qubit.

5.2.3 DNA Sequencing

Sequencing was performed by the CGR at the University of Liverpool using Novaseq SP 2x150bp. One lane on Illumina Novaseq using SP chemistry (paired-end 2x150bp sequencing), generated an estimated 325 million clusters per lane (ENA study PRJEB39463).

5.2.4 Sequence Read Processing

The raw fastq files were trimmed using using cutadapt version 1.2.1 [300] to remove Illumina adaptor sequences. Option -o 3 was used so that the 3' ends of any reads that match the adaptor sequence for 3bp or more are trimmed.
The reads were trimmed using Sickle version 1.200 [301] with a minimum window quality score of 20. Reads that were shorter than 15bp after trimming were removed. If only 1 read pair passed this filter it is included in the R0 file.

5.2.5 Variant Calling

The Genome Analysis Toolkit (GATK) was used to discover variants between the genomes and to mark the SNPs [453] [454]. The pipeline was used according the GATK best practices. First, fastQC [455] was used to assess the quality of the sequence files, both R1 and R2. The reads were mapped to my *T. foetus* annotated genome using genome using BWA (version 0.7.17 r1188) [305] [456] which created a SAMfile using samtools (version 1.9) [305]. The resultant SAM file was converted to a BAM file also using samtools [305].

Picard [457] was used for pre-processing of the BAM files before they were passed to the GATK pipline [453]. The steps used were: AddorReplaceReadGroups, CleanSam, FixMateInformation, MarkDuplicates, IndelRealligner, CallableLoci and Haplotype caller.

AddorReplaceReadGroups assigns a unique identifier to each BAM file. This means that identified SNPs can be assigned to a specific BAM file or sample.

CleanSam cleans the provided BAM file, sets MAPQ to 0 for unmapped reads and performs soft-clipping beyond the end of the reference alignment.

'FixMateInformation' ensures that all mate-pair information is matching between both pairs.

Duplicates were marked using 'MarkDuplicates'.The duplicates come from where the same piece of DNA has been sampled more than once. This can cause errors with finding true variants. BWA can sometimes 'clip' the end of sequence when it has mapped, therefore some fragments, i.e. those mapped to the reverse strand, can be identified by 3' instead of 5'. These duplicates can cause overrepresentation in the sequence quality, therefore, they must be marked and removed from the samples.

'IndelRealligner' was used to reduce the number of mismatched bases between the BAM file sequence and the reference genome sequence. It finds the best alternate consensus sequence caused by the the indels.

'CallableLoci' determines which loci can be called and the Haplotype Caller was used to identify SNPs and indels within the BAM files and were called separately for each sample. Variants from each sample were joined using the Haplotype Caller and then saved in a variant call format (VCF) file. The VCF file contains information, such variant position and quality. The GVCF files of all samples were merged together.

5.2.6 Filtering of SNPS

Just the SNPs were selected using 'selectvariants'. Those that passed or failed the quality criteria were tagged using 'variantfiltration'. Any that lack the PASS annotation were filtered out by 'selectvariants'. The SNPS were then filtered to genotypes using 'vcftools'. This produced a list of all contigs and whether the SNPs possess:
1) homozygous reference i.e two copies of reference allele (0/0).
2) heterozygous for reference genome i.e. one allele is the same as the reference genome and another is different (0/1).
3)homozygous for non-reference allele i.e. two copies that are not the same as the reference genome (1/1).

The gene list was processed through a perl pipeline to extract the SNPS and a file was created with all the SNPs in each gene, their position, the nucleotide change and whether it is a synonymous or non-synonymous mutation.

5.2.7 Per Gene Population Statistics

Genome statistics were calculated using popgenome [458] [459]. This programme was used as it can handle whole genome data. The SNPs that were identified in introns and untranslated regions were filtered out, as were genes that did not contain SNPs. A list was produced of all the SNPs present in the exons from the *T. foetus* strains. Popgenome was used to calculate statistics: the nucleotide diversity and Tajima's D values for each protein-coding gene based on the SNPs.

5.3 Results

5.3.1 Trichomonad DNA Reads

DNA from six trichomonad strains were sequenced by the University of Liverpool CGR. After trimming there was a range of 108-156 million trimmed reads produced for the different samples (Table 5.2).

Strain	No. Raw Reads	No. Trimmed reads
Belfast	113,680,106	113,309,480
KV-1	116,656,664	116,288,048
DK-2	114,995,714	114,636,161
BP-4	157,339,356	156,774,360
UT	123,295,706	122,859,639
T. mobilensis	108,896,432	108,342,358

Table 5.2: Numbers of raw and trimmed reads of Illumina sequenced DNA of multiple trichomonad strains

5.3.2 Genomic Variation

SNPs were found in 2,543 contigs in the genome when all samples were included: all five *T. foetus* strains and *T. mobilensis*. In the *T. foetus* samples only there were 102,986 SNPs found over 2,461 contigs. When the number of SNPs per kb was calculated there was a range of 0-72 SNPs per kb across the genome (Figure 5.2). 60,094 kbs along the genome contained no SNPs, with 26,264 containing 1 and 11,043 containing 2. As the number of SNPs per kb increased, the frequency of kbs containing that number decreased. Only one kb contained 72 SNPs. There were also three other kbs that contained 50 or more SNPs.

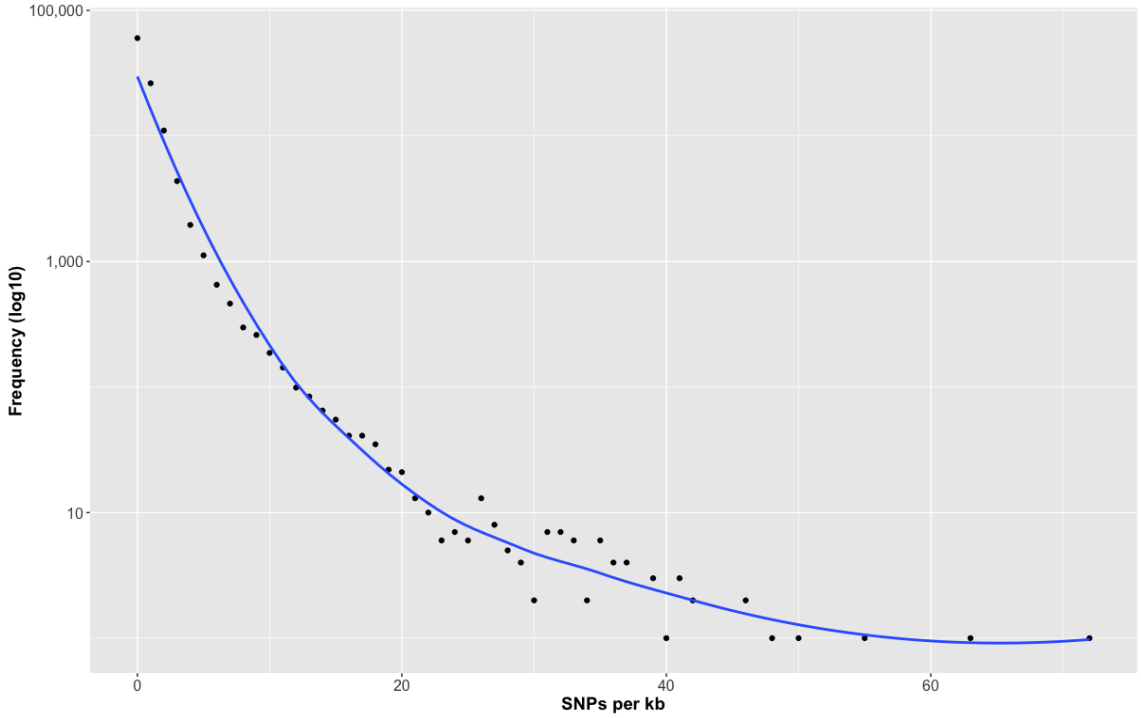

Figure 5.2: The frequency of SNP number per kb of the *T. foetus* genome. Five strains of *T. foetus* were used in the analysis and there is a range from 0 to 72 SNPs per kb with the majority of kbs along the genome possessing 0 SNPS. A trendline was also added.

Strain	Heterozygous SNPs	Homozygous SNPS
Belfast	9,009	1,440
DK-2	15,576	24,930
UT	23,037	52,328
BP-4	8,806	1,435
KV-1	22,826	52,441
T. mobilensis	787,776	33,928

Table 5.3: The number of homozygous and heterozygous SNPs found in different strains of *T. foetus* and *T. mobliensis* when compared to the reference *T. foetus* Belfast strain genome annotation.

Between all of the trichomonad samples there are very large differences in the number of SNPs found (Table 5.3). For instance, *T. mobilensis* was found to possess 787,776 heterozygous SNPs, meaning one copy was the same as the genome reference and one copy was different and 33,928 homozygous SNPs where both copies possessed a different nucleotide to the reference. This is over 30 times higher than any of the *T. foetus* samples.

Within the *T. foetus* samples, there was, again, a large range of SNP number, with the Belfast and BP-4 strains having 1,440 and 1,435 homozygous SNPs respectively and the UT and KV-1 strains having over 52,000 homozygous SNPs. The number of SNPs were relative to the original Belfast strain reference genome, produced by PacBio sequencing.

5.3.3 SNPs in Gene List

When all the trichomonad strains were used in the SNP analysis, five *T. foetus* and one *T. mobilensis*, in total 2,023 SNPs were identified in the antigen shortlist (77 genes). There were 1,126 non-synonymous mutations and 897 synonymous mutations found. There was a large range of SNP number between the different genes, with 2 SNPs identified in TTF14449 and 265 SNPs identified in TTF45042 (Table D1). Every gene contained at least one SNP and there were only two genes that did not contain any non-synonymous mutations, TTF00576 and TTF08916.

When only the *T. foetus* strains are compared with the reference genome, the number of SNPs identified per gene are significantly lower, with the highest number being 35 in TTF04635. In total there were 174 SNPs found; 133 non-synonymous and 41 synonymous (Table D1). There are 23 genes that do not contain any SNPs.

Across the shortlist there was a range in d_N/d_S ratio with 31 of the *T. foetus* only genes having a ratio of 1 or 0 and one gene (TTF75601) having a ratio of 8.

5.3.4 Whole Genome Statistics

Across the 21,053 genes of the *T. foetus* samples that contained SNPS, the Tajima's D values were spread from -2 to 3. There were 18,888 genes that were identified as having a positive Taijima's D value and 2,161 that had a negative value. Those with a value of 0 had been filtered out of the analysis in a previous step due to not possessing any SNPs.

As 21,000 genes from the strains out of a total of 84,000 genes in the *T. foetus* genome contained SNPs, one quarter of the genome appears to be under selection pressures.

There is a weak positive correlation between the Tajima's D values and nucleotide diversity (Figure 5.3). The overall majority (66,600) of genes in the strain genomes have a nucleotide diversity of 0 as they have been filtered out due to not possessing SNPs and 13,396 genes have a nucleotide diversity between 0 and 1. However, four genes (TTF25397, TTF03299, TTF33699 and TTF77035) have a

nucleotide diversity of over 40 in terms of mutated sites. Overall, the majority of genes (75%) in the *T. foetus* genome appear to have a nucleotide diversity of 0 and 90% have one changed site or less.

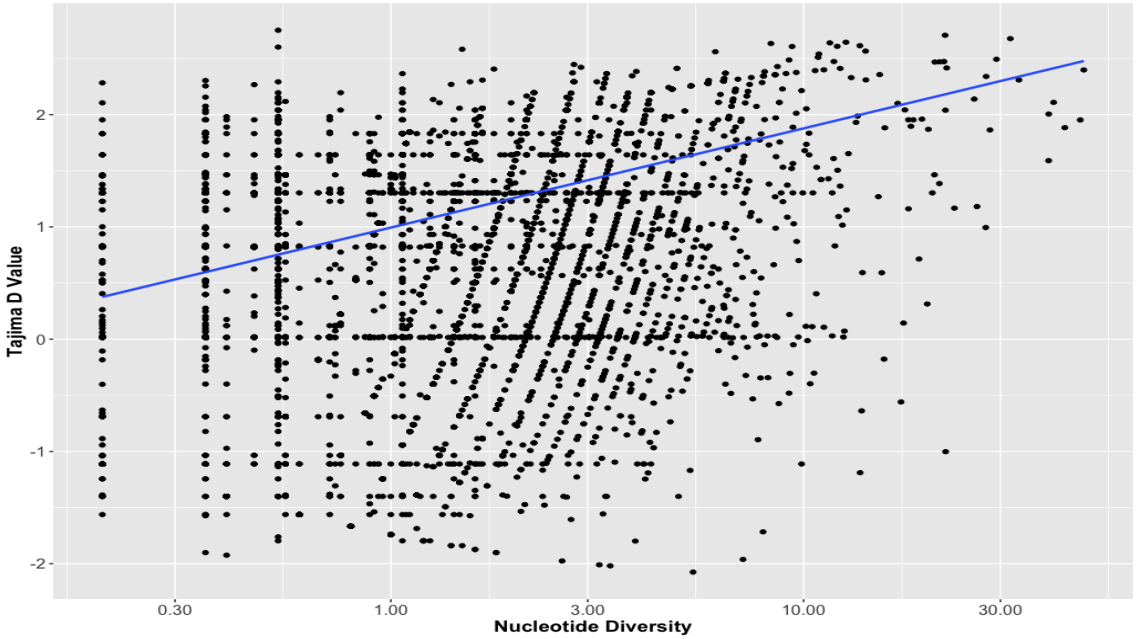

Figure 5.3: Tajima's D value compared to log10 nucleotide diversity, in terms of number of changes, of genes containing SNPs between the *T. foetus* samples and the reference *T. foetus* genome. SNPs between only the different *T. foetus* strains and the reference genome were taken into account. Both the Tajima's D value and nucleotide diversity were calculated using Popgenome [458][459] in R.

5.3.5 Gene Shortlist

The genes in the shortlist were given a score according to certain criteria. The total number of SNPs, both synonymous and non-synonymous; the fold change was the expression of each gene in the presence of a cell monolayer relative to controls grown in Diamond media and peptide immunogenicity is a measure of all of the immunogenic epitopes seen in both the natural and experimental infections as identified by the peptide array. The scores for each criterion were calculated and added together. The genes were ranked based on their total score and the top 8 were selected (Table 5.4).

Gene ID	Product	Total no. of SNPs	Gene Expression FC	Total No. Immunogenic peptides
TTF53402	hypothetical protein TRFO_12669	0	1.88	163
TTF00910	hypothetical protein TRFO_06137	0	1.85	151
TTF05727	hypothetical protein TRFO_04170	0	10.31	0
TTF73824	hypothetical protein TRFO_29759	1	1.39	91
TTF72511	hypothetical protein TRFO_29759	5	5.59	90
TTF08916	hypothetical protein TRFO_04830	0	8.81	0
TTF48521	hypothetical protein TRFO_11786	0	7.13	0

Table 5.4: Vaccine candidate gene shortlist. All genes fulfil the vaccine candidate criteria from previous chapters including the presence of a transmembrane domain. They contain few or no non-synonymous or synonymous mutations in any *T. foetus* strain examined relative to the *T. foetus* reference genome. They are also preferentially expressed in the presence of a cell monolayer relative to controls and for those that were used in the peptide array, they have a high combined total of immunogenic peptides from both experimental and natural infections.

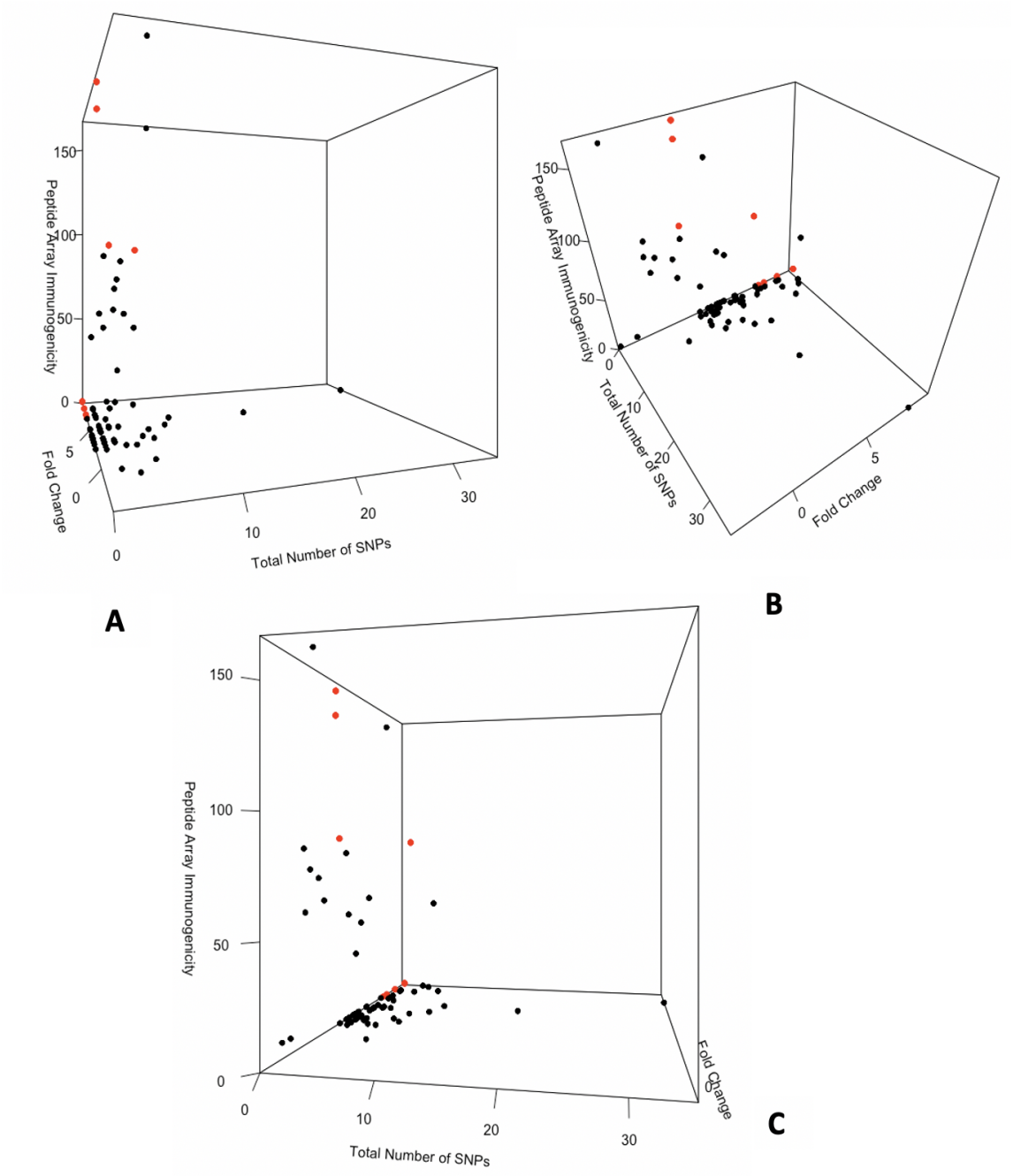

Figure 5.4: 3D plots, shown from different angles, (A, B and C), created in R showing the SNP number, immunogenicity and fold change of the vaccine candidate shortlist. The total number of SNPs includes both synonymous and non-synonymous mutations, the fold change was the expression of each gene in the presence of a cell monolayer relative to controls grown in Diamond media. The peptide immunogenicity is a measure of all of the immunogenic epitopes seen in both the natural and experimental infections as identified by the peptide array. The 8 points highlighted in red are those in the final gene shortlist as, when all genes were ranked according to all three criteria, these had the highest final score and were therefore, most suitable for taking forward for further testing.

5.4 Discussion

This chapter aimed to identify the SNPs in five *T. foetus* strains relative to the assembled and annotated Belfast strain genome and to see whether SNPs were present in the list of possible vaccine candidates from Chapter 4. Overall, the short list of 77 candidates has been reduced to 8 well conserved antigen genes (Table 5.4).

Commonly, SNPs can produce issues for designing vaccines, for example, the *Plasmodium falciparum* circumsporozoite protein (PfCSP) [460] is a well studied, potential vaccine candidate for malaria. However, genetic polymorphisms, such as SNPs could be an issue regarding the efficacy of the vaccine. In Myanmar, 51 full length genes were amplified from blood samples from 51 patients and polymorphic diversity was analysed. Polymorphic diversity was also analysed from other, publicly available PfCSP genes as a comparison [460]. Low genetic diversity was seen in the global population but there were geographic differences. The central repeat region of the Myanmar PfCSP was found to be highly polymorphic, whereas the N and C-terminal repeat regions were conserved. The differentiation between geographic isolates showed evidence of recombination which would need to be taken into account for a vaccine.

Saralamba *et al.* (2018) [461] also analysed blood samples from 89 Malaria patients from various Asian countries in addition to 58 previously published *Plasmodium falciparum* sequences from Africa. There was clear clustering of isolates by geographic origin with high genetic differentiation between the Asian and African countries [461]. They also identified fewer genetic repeats in the Asian samples but these also showed higher nucleotide diversity [461].

5.4.1 Whole Genome SNPs

Across the whole genome, there were more genes with a positive (18,888) Tadjima's D value than negative (2,161) suggesting a potential lack of rare alleles or the genes are undergoing balancing selection. Additionally, the nucleotide diversity, in terms of mutated sites, across 90% of the genome was either 0 or less than 1 suggesting that the majority of genes are not undergoing mutations or strong selection pressures.

All strains used were laboratory strains, not natural infections, and had been obtained from North America and Europe. In order to identify whether a vaccine cocktail would be required, it would be beneficial to analyse SNPs from *T. foetus* strains found in different continents, such as Africa or Asia.

5.4.2 SNP Number Between Strains

There was a much larger number of heterozygous SNPs found in the *T. mobilensis* sample compared to the *T. foetus* samples. This is due to the fact that *T. mobilensis* is less genetically similar to the reference genome which was created from the Belfast *T. foetus* strain. The re-sequenced Belfast strain has a low number of homozygous SNPs compared to *T. mobilensis* and the other *T. foetus* strains, apart from strain BP-4. This low number, 1,440 is to be expected as the original reference genome was created from the Belfast strain. It is, therefore, surprising that BP-4 also has a similar number of homozygous SNPs, 1,435 and an even lower number of heterozygous SNPs (8,806) compared to the Belfast 9,009 (Table 5.3). In *Giardia* [462] there were over 100,000 SNPs found so this number seems inkeeping if low.

The SNPs seen between the original PacBio reference and the Illumina re-sequenced Belfast sample may be due to the differences in genome sequencing, especially as it is known that PacBio sequencing can have a relatively high error rate [224]. Additionally several years had passed between the original sequencing and the re-sequencing and so changes to the Belfast strain could have occurred in the lab.

There do not appear to be clear similarities in number from strains from the same geographic areas, for instance, the two *T. foetus* strains that were isolated in Beltsville, Maryland (BP-4 and UT) have very different SNP numbers. There were 8,806 heterozygous SNPs identified for BP-4 and 23,037 for UT and 1,435 homozygous SNPs identified for BP-4 and 52,328 for UT (Table5.3). The number of SNPs between all of the American samples seems highly variable. There may be a large amount of variation between strains, in a similar way to malaria in Africa.

However, when the number of SNPs between the feline strain of *T. foetus* and the bovine and porcine strains were compared, only 68 SNP heterozygous variants were found between the bovine and porcine samples [120], much lower than any comparison seen in my analysis (Table 5.3). When the feline strain was compared to the bovine and porcine strains, 65,569 and 65,615 SNPs were identified respectively [120]. This suggests that the different geographic regions and time points when the samples were collected appear to have a large effect on the genetic distinctness of the strains.

All the *T. foetus* strains used in these experiments were laboratory strains, first isolated over 50 years ago. The adaption of these strains to lab culture and freeze-thawing of samples could have

lead to a change in genetic diversity, such as bottle-necking or an overall general loss of diversity. Comparing the findings here to freshly isolated strains may further genetic differences.

5.4.3 Shortlist of Vaccine Candidates

The genes chosen as the vaccine candidate shortlist have no or very few synonymous or non-synonymous mutations between themselves and the reference genome, suggesting that these genes are not subjected to high selection pressures and do not commonly mutate. In this way, they are more likely to be structurally invariant. Additionally, they fulfil the criteria of the previous chapters of containing a transmembrane helix. They are also all preferentially expressed in the presence of a host monolayer. For those that were present in the peptide array, all show a high number of immunogenic epitopes. Furthermore, TTF08916 has no non-synonymous mutations when *T. mobilensis* is used in the analysis, so there may be the possibility of a cross species vaccine candidate.

As many of these genes do not contain either synonymous or non-synonymous SNP mutations they were filtered out at an earlier stage. Due to the lack of mutations, their Taijima's D value, nucleotide diversity and the d_N/d_S ratio will be 1. These genes have been found to contain transmembrane motifs, are have increased expression in the presence of the host and appear to be structurally invariant across the populations.

Genes were looked at using Phyre, InterProScan [281] [282] and BlastP [278]. Looking at the amino acid composition, all vaccine candidate genes have more serine (S), threonine (T) and asparagine (N) residues than would be expected on average. Serine, threonine and asparagine are routinely sites for glycosylation [463] so it is highly likely that the genes are heavily glycosylated and decorated with sugars. This is usually indicative of cell-cell interactions [464].

1) TTF53402 is 1560 amino acids in length and has a combined serine, threonin and asparagine (S+T+N) value of 21.7%. It also was found to have an interpro [281] [282] match to a bacterial invasin protein (IPR008964) and has a homolog in *T. vaginalis* (XP_001329076.1). There was also a Phyre match to a sugar binding protein (confidence 99%, coverage 9%). This suggests that this gene is a glycoprotein but not a member of any recognised family.

2) TTF00910 is 1764 amino acids in length and has an S+T+N of 25.1%. InterProScan showed no domains, however, Phyre [465] shows a match to a viral glycoprotein (confidence 86.8%, coverage 3%).

3) TTF05727 is 236 amino acids in length and has an S+T+N of 19.9%. like TTF53402 it has a homolog in *T. vaginalis* (XP_001302612.1), however, like TTF00910 InterProScan showed no domains. There is also a Phyre match to a member of e2f family of transcription factors (100% confidence, 73% coverage).

4) TTF73824 is 943 amino acids in length and has an S+T+N of 25.3%. No homologs or domains were identified but there was a Phyre match to an alginate lyase (confidence 87%, coverage 54%). This may be a similar enzyme which is involved in the breakdown of host proteins.

5) TTF38179 is 1068 amino acids in length and an S+T+N of 25.5% and contains a PT-domain found in bacteria and phytophthora and may fulfil a similar function.

6) TTF72511 is 153 amino acids in length and an S+T+N of 26.1% and contains no homologs or domains.

7) TTF08916 is 153 amino acids in length and an S+T+N of 19.0%. It has a homolog in *T. vaginalis* (XP_001314982.1)but no matches with InterProScan. Like TTF05727 it has a Phyre match to an e2f family of transcription factors (100% confidence, 89% coverage).

8) TTF48521 is 529 amino acids in length and an S+T+N of 25.5% and like TTF08916, TTF53402 and TTF05727 it has a homolog in *T. vaginalis* (XP_001320645.1). Though there were no InterPro domains, Phyre matched to an alginate lyase, in a similar way to TTF73824 (96.8% confidence, 23% coverage).

The genes in the shortlist suggest that they have some capacity for cell-cell interactions, namely in their high serine, threonine and asparagine content and the likelihood that they are highly glycosylated. As three have homologs in *T. vaginalis*, they appear conserved, another advantage for a vaccine candidate.

5.4.4 Conclusion

In this chapter, SNPs between the five *T. foetus* strains and one *T. mobilensis* strain have been identified. These SNPs, both synonymous and non-synonymous, have been crosschecked with the vaccine candidate gene list produced in Chapter 4 in order to identify which, if any, candidates are structurally invariant across populations (Figure 5.5). Eight genes were identified from the shortlist of 77 genes that possessed few, if any SNPS, were upregulated in the presence of the host and many showed strong immunogenic responses in the peptide array. These genes can then be tested to see

if it is possible to express them in other organisms, such as yeast and preliminary clinical trials could be performed.

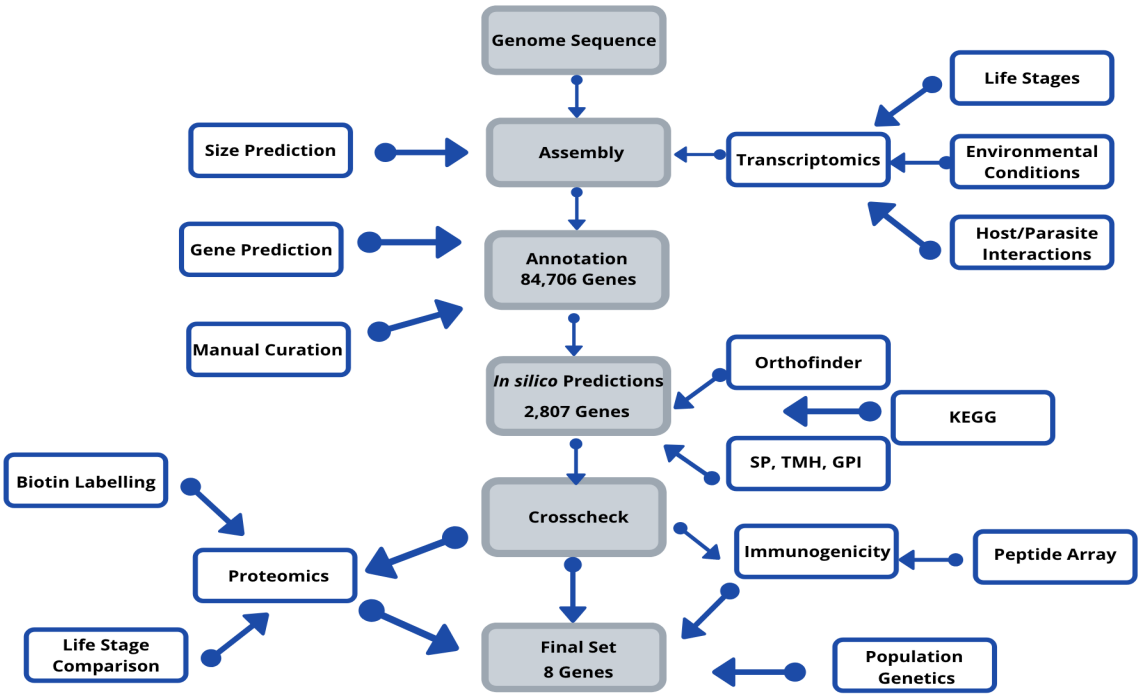

Figure 5.5: Flow diagram showing the reduction of genes from the *T. foetus* genome from the initial annotation (87,000 genes) to a final set of 8 genes. The genes were tested for their structure, presence of transmembrane helices, immunogenicity and upregulation in the presence of the host. They also appear to be structurally invariant across the populations tested.

Chapter 6

General Discussion

This thesis describes the identification of vaccine candidate antigens against the bovine parasite *Tritrichomonas foetus*. Starting from the assembly and annotation of the *T. foetus* genome to identifying a small subset of putative cell surface antigens as potential vaccine candidates.

The *T. foetus* genome was assembled using a variety of methods to produce a final genome size of 147Mb. Open reading frames were identified and manually curated. This yielded a gene set of 84,756 genes. A network was produced containing the entire predicted *T. foetus* cell surface predicted proteome, showing gene families from *T. foetus* and *T. vaginalis*. The proteins included had key features that would make them likely to be cell-surface expressed. Fifty proteins were selected from the network. These proteins were used in a peptide array to test the immunogenicity of the epitopes of the peptides produced. This produced a list of peptides with the highest number of immunogenic epitopes and high maximum intensity which could be highly credible vaccine candidates.

Several strains of *T. foetus* and *T. mobilensis* were grown in co-culture with a confluent MDCK monolayer and transcriptomes were produced. This was to identify genes that were expressed in the presence of the host cells. The genes that were found to be preferentially expressed in the presence of the host cells were cross-checked with the peptide array proteins to produce a list of 77 candidate antigens. These candidate antigens were examined for SNPs to see if they were structurally invariant. The genes were then cross-checked, comparing the number of immunogenic epitopes, transcriptomic fold change in the presence of a host monolayer and number of SNPs relative to the reference genome. The genes were ranked for each of these attributes and the total scores across the three criteria were calculated. In total, the 8 genes with the highest scores were selected.

6.1 Future Steps in Vaccine Development

As stated by Rappuoli *et al.* (2016) [197] and Dalsass *et al.* (2019) [466] one of the next steps in determining the most successful vaccine candidates is the expression of the proteins. The proteins from the genes in the shortlist could be expressed *in vitro* either in *Escherichia.coli* or a yeast such as *S. cerevisiae*. This is particularly important as the proteins need to be expressed with the correct post-translational modifications (PTMs), such as glycosylation. The cell-surface localisation of these proteins also has to be confirmed experimentally and B and T cell epitopes could be predicted bioinformatically [197]. Furthermore, trials could be performed using either individual proteins, or a cocktail of proteins, first in mice and then in cattle to test for their protective effect in challenge experiments. A range of adjuvants would also have to be trialled and tested to see which gives the highest immunological effect [197] [466] and so can go forward to preliminary clinical trials.

6.1.1 SNPs between strains

It would be interesting to know whether there is an increase in SNPs between strains of *T. foetus* that infect different hosts, such as comparisons between feline, bovine and porcine strains. All strains examined in this thesis were originally isolated from infected cattle and show a large range in SNP number, so there is likely to be even greater differences between strains from different host species. Furthermore, if some of the 8 vaccine candidates identified were also found to not possess SNPs in the feline isolated strains, there may be potential for a vaccine that can protect multiple host species, although there are known to be large numbers of SNPs between the feline and bovine *T. foetus* strain [467]. However, if vaccine candidates can be identified between five *T. foetus* strains and *T. mobilensis*, a related but distinct species, it is likely that genes containing no SNPs between the different *T. foetus* genotypes can be found.

Additionally, as two genes within the 77 candidate antigen shortlist were found that did not possess any SNPs when all five *T. foetus* and the *T. mobilensis* samples were compared with the reference, it may be possible to find genes from other trichomonad species that are conserved, such as *T. gallinae* or *T. vaginalis*. This may lead to a trichomonad vaccine.

6.1.2 Recombinant expression

One of the key criteria for a vaccine candidate is its ability to be expressed recombinantly so that it can be produced in large quantities. Producing a recombinant vaccine protein involves transfecting

cells with the gene of interest inserted into a vector. A protein can then be produced by the host cell and this can be harvested, purified and used.

Both prokaryotic and eukaryotic expression systems are widely used [468].
Bacterial expression systems are useful because they are relatively cheap, easily scalable and easy to mass produce proteins [468]. Zhang *et al.* (2020) [469] [470], for example, used *E. coli* to express proteins against *Toxoplasma* and *T. vaginalis*. However, there can be issue when expressing mammalian proteins and they can form inclusion bodies and also do not have the ability to add post translational modifications (PTMs), such as glycosylation [468]. *E. coli* and other bacteria can also produce low yields and inclusion bodies [471].

Yeast can express eukaryotic proteins and also has the advantage of being able to express secreted proteins as well as intracellular proteins [468]. It can also produce both O-linked and N-linked glycosylation, both of which are found in *T. foetus* [414]. The yeast *Pichia pastoris* has been found to have advantages over *S. cerevisiae* [471]. *S. cerevisiae* contains sugars, for example, mannose [472] that can affect glycosylation and so these sugars need to be removed before the recombinant protein is produced. *S. cerevisiae* can also hyper-glycosylate the protein by adding hundreds of mannose residues. *K. phaffii* or *Pichia pastoris* comparatively, adds fewer residues, which can somewhat overcome the problem.

Glycosylation and glycosylation profiles are important for several reasons. Certain antibodies can recognise glycosylation [473] and can also be blocked by glycosylation in high levels [474] and can prevent the protease action of antigens [474]. Key proteins involved in antigen recognition, such as the Major Histocompatibility Complex (MHC) are glycosylated [474] and the patterns of which can also affect MHC uptake. MHCs can present glycopeptides and this may affect T-cell recognition.

Mammalian cell culture systems also allow for glycosylation but do not have the potential drawback of over-glycosylation of proteins, however, it is much more expensive [468]. Additionally, different mammalian cell lines have different glycosylation effects, e.g CHO or HEK cells [475] so this has to be taken into account when choosing cell lines. CHO cells, for example, do not express certain glycosylation enzymes (Gal α-2,6 ST, and α-3/4 fucosyltransferase) which, in human cells at least, is required for certain glycostructures and human-glycoproteins. Additionally, when glycans produced by both CHO and HEK cell lines were compared, HEK cells had more complicated glycosylation profiles [475]. Specific yields from mammalian cells lines can be very low and expensive to produce [471].

Pichia pastoris was also deemed as more useful than *E. coli* and *S. cerevisiae* by Morton *et al.*

[476] (2000). When rabbit liver carboxylesterase was expressed in *E. coli*, *Pichia pastoris* and *S. cerevisiae* different amounts of enzymatic activity were seen in the recombinant proteins. Very little enzymatic activity was identified in the recombinant protein produced by *E. coli* and little by *S. cerevisiae*, however, active protein was produced using *Pichia pastoris*.

As *T. foetus* appears to use both O-linked and N-linked glycosylation [414], either the yeast or mammalian cell culture systems would appear to be the best method for protein expression. The vaccine candidate gene would need to be able to be expressed in some form so that it could be mass produced and used in clinical trials

6.1.3 Serum Lysis Assays

In order to test for the protective effect of the candidates and whether antibodies or complement have been produced, several methods have been developed. Serum from animals post infection is extracted and added to *T. foetus* cells to see whether they induce killing of the parasites. Zhang *et al.* [469] [470] collected antisera from both rats and mice that had been injected with Freund's adjuvant and recombinant proteins, either TvAP33 for the *T. vaginalis* experiments or TgTH for the *Toxoplasma* experiments. This antisera was then used in later expression and localisation experiments.

Meulenbroek *et al.* (2014) [477] used non-quantitative haemolytic assays to assess the lysis capabilities of serum from patients on bromelain-treated red blood cells (RBCs). Fluorescently labelled cells were used in combination with FACS for spectrophotometric detection of lysate, for example, haemoglobin in the case of red blood cells. Whereas, Stasilojc *et al.* (2020) [478] used calcein release to measure serum complement activity and Costabile *et al.* (2010) [479] used a CD50 assay to measure the complement response induced. This CD50 assay tested the capability of serum from sheep to lyse sheep red blood cells that are coated with haemolysin. A similar assay could be used to show how strongly this pathway is induced in cattle that have been infected with *T. foetus*.

Weerts *et al.* (2019) [480] used a bacteriocidal shigella assay to measure the antibody killing ability and activating complement response in serum from rabbits, i.e. where there is attachment and lysis of the bacteria (or parasite) by the antibodies found in the serum. This also has the advantage that it can be performed in 96-well plates so it has high throughput abilities and could also be used to quantify antibody titres.

Another way of measuring antibody production and damage in the sera is using the Sabin-Feldman

dye test. Dando *et al.* (2001) used this method [481] to identify whether antibodies were present in sera of clinical patients against *Toxoplasma*. If antibodies are present, they prevent methylene blue from entering the *Toxoplasma* cells. If cells do stain blue, however, there are no antibodies. This provides an easy to quantify result. One of these methods, or a combination of methods could be used to identify whether infected cattle have produced functional antibodies against *T. foetus* infection.

6.1.4 Expression localisation

It has been shown that the vaccine candidate genes all possess a transmembrane helix that suggests that the protein is located on the cell surface based on its primary and secondary structures. However, the cell-surface localisation of the proteins would need to also be confirmed experimentally.

Zhang *et al.* (2020) [469] used cell-surface localisation for both *Toxoplasma* [469] and *T. vaginalis* [470]. *Toxoplasma* cells were fixed on slides, washed using rat antiserum, PBS and then goat anti-rat IgG antibody with Cy3 labelling. This was in order to fluorescently label the proteins of interest. DAPI was added to identify the nucleus, and cells were examined using fluorescence microscopy. Meissner *et al.* (2011) [482] similarly used fluorescent tags and microscopy to determine cell-surface localisation. They used GFP tags with *C. elegans* then fluorescently imaged the cells to identify localisation of body wall muscle (sarcomeric) genes.

Molecular sensors can now also be used [483] by way of the co-localisation of fluorescent markers, however this is low throughput. Fitzgerald *et al.* (2020) [483] have developed a SNAP-tag substrate that can be conjugated to the protein of interest and the trafficking of said protein can be detected using flow cytometry. This allows for quantitative tracking of the protein localisation. The SNAP-$^{\text{switch}}$ is a localisation sensor. When the SNAP-$^{\text{switch}}$ interacts with the SNAP-tag a quencher molecule is transferred to the SNAP-tag and a Cy3 fluorophore becomes permanently attached to the protein of interest. The fluorescence can then be measured and tracked [483].

Cell surface localisation can also be determined biochemically. Besingi *et al.* (2015) [484] used susceptibility to extracellular protease digestion as a way of identifying if a protein of interest is on the cell surface. An intracellular reporter protein is attached (in bacteria) that is undigested if the membrane remains intact but is digested when cells are lysed. The intracellular reporter will have access to the extracellular protease if the protein it is attached to is on the cell surface but, if

it remains within the cell, the protein will not be digested. These techniques could provide further evidence that the 8 *T. foetus* candidate antigens are indeed cell-surface expressed.

6.1.5 Challenge Models

Challenge experiments are needed to show that the vaccine candidates confer protection against the pathogen or disease. Mice or cows are given the vaccine and then challenged by infection with the parasite to see if there is a reduction in severity or length of the infection or no infection at all.

In experiments performed by Zhang *et al.* (2020) [470] on the *T. vaginalis* adhesion protein, 80 Balb/c mice were used. They were split into four groups of 20 mice and were injected with the *T. vaginalis* putative recombinant adhesion protein TvAP33 with an adjuvant. They were vaccinated 3 times with the protein on day 0,14 and 28. On day 38 the mice were challenged using $1x10^7$ *T. vaginalis* parasites any mice that showed clinical signs in the following month were euthanised. The survival rate of the mice was significantly higher when the protein was used than in the controls, with 100% survival 20 days after the challenge. In order to identify whether immunity is long lasting when *T. foetus* antigens are used, further challenge experiments would be needed, at least a month, two months and six months. this would show whether there is immune memory and the initial results seen are not only the immediate vaccine response. Current *T. foetus* vaccine attempts provide only short term immunity and so multiple vaccinations and boosters are needed which is not practical in large herds.

Zhang *et al.* (2020) [469] also performed a similar experiment with *Toxoplasma*. Mice were again immunised 3 times and challenged 10 days after last inoculation. The survival rate was measured-after 30 days post challenge, the brains of all mice were removed and analysed to identify the number and size of *Toxoplasma* cysts present. It would seem that a similar experiment could be performed using *T. foetus* proteins once they have been shown to be recombinantly expressed. Ideally the vaccine candidates would induce $CD4^+$ T cells as these play a key role in B cell proliferation and switching [197]. $CD4^+$ effector activity is also commonly needed against parasite infections [197]. The antigens could be trialled both individually and in combination to identify which has the most protective effect against *T. foetus.*

Challenge experiments are also commonly performed in cattle as there are clear physical differences between mice and cattle, these would prove to be a more accurate experimental model for a *T.*

foetus vaccine. Edrington *et al.* (2013) [485], for example, used a challenge model of *Salmonella* in the lymph nodes of cattle. Calves were vaccinated with a commercially available *Salmonella* vaccine and then challenged with *Salmonella* after 14 days. The *Salmonella* was recovered from the lymph nodes of fewer vaccinated calves than controls and a modest vaccine effect was identified. Albanese *et al.* (2018) [486] used a similar method with chickens. They used chickens to test a *Coccoidia* vaccine with readings were taken 7 days post challenge to evaluate acquired protection. In this case the challenge experiment showed that the vaccine was successful.

Cooper *et al.* (2018) [487] used human challenge models for evaluating malaria vaccine candidates, they were also able to test multiple strains. Multiple testing in this way facilitates pre-clinical development and can allow for faster clinical trials, particularly as multi-strain vaccines can be more relevant for immunity. Challenge experiments using candidates found in multiple strains of *T. foetus* could lead to faster and more relevant clinical trials in cattle.

6.1.6 Mouse and cattle trials

Matuschewski *et al.* (2012) [488] tested their vaccine for malaria using murine infection models. All vaccines require clinical trials before they can be licensed. Mouse models allow for larger scale pre-clinical experiments than using cattle would allow due to their small size and relatively low cost comparatively to cows. In malaria, a vaccine candidate that was thought to be successful in phase II trials, SPf66, was later found out to not have a significant protective effect in phase III trials [488]. As there was no mouse model for this candidate it led to great difficulties to make improvements to the vaccine design.

After experiments in mice are performed, there would also need to be experiments performed in cattle, as these are the species the vaccine is designed for and because there are clear physiological differences between mice and cows. They may have different immune responses or susceptibilities that are not shown through mouse trials alone.

When Buddle *et al.* (2013) [489] were testing an oral TB-BCG vaccine against *M. bovis* they identified parameters that affect vaccine efficacy in the cattle tested. These included: age of the cow, as calves less than 1 month had the best efficacy, dose, pre-exposure to *M. bovis* and the strain of BCG used. Similar criteria would have to be taken into consideration for a *T. foetus* vaccine and trials, particularly the age of the cattle and strain of *T. foetus*. Duration of infection and timing of booster vaccinations would also need to be considered. As mentioned in the introduction, older

bulls are usually more affected by *T. foetus* than younger bulls [25]. It has also been shown in this thesis that different strains of *T. foetus* have different levels of cytotoxicity and so may show different effects in trials or challenge experiments.

When Farooq *et al.* (2019) [490] were testing a vaccine for haemorrhagic septicaemia, they had to test whether it was safe in multiple types of cattle. This is also relevant for a *T. foetus* vaccine as there are many breeds of cows and they can be affected differently by *T. foetus*. As mentioned in the introduction certain species and breeds of cow are more likely to retain *T. foetus* infection than others [26]. For example, *Bos taurus* are more likely to retain infection than *Bos indicus*, particularly the Angus, Simmental and Charolais breeds [26]. Testing the vaccine candidate in multiple breed and species of cattle could lead to clearer picture of how effective the vaccine is, particularly as natural infections could show greater variation than laboratory infections.

6.2 Conclusion

In this thesis, a *T. foetus* genome has been sequenced and assembled, and annotated using diverse transcriptomic evidence. The basis for the selection of candidate antigens depended on a set of assumptions set out in the general introduction. Assumptions such as these are necessary, particularly when the identity of proteins is unknown and there are many thousands to reduce the diversity to a manageable list of candidates. These assumptions are based on rational criteria but may not always be accurate, meaning that some potentially useful antigens are ignored. From a total number of 84,706 protein-coding genes, 2,807 were identified to be cell surface expressed and parasite specific using *in silico* methods. Of these, 77 genes were identified from proteomic and transcriptomic assays to be both preferentially expressed in the presence of host cells and also among the most immunogenic antigens in natural and experimental infections. Population genetic approaches were used to examine the predicted antigens for structural variation amongst parasite strains. Finally, 8 antigens have been identified as the most likely *T. foetus* vaccine candidates because they have highly conserved protein structures across the parasite population. These candidates are now available to move to the next stages in the reverse vaccinology pipeline, such as recombinant expression and challenge experiments, and may provide a way towards eradicating bovine trichomoniasis in livestock across the globe.

Bibliography

[1] D. Hudson, L. Ball, J. Cheney, R. Mortimer, R. Bowen, D. Marsh, and R. Peetz, "Development and Testing of a Bovine Trichomoniasis Vaccine," *Theriogenology*, vol. 39, no. 4, pp. 929–935, 1993.

[2] M. Benchimol, "Trichomonads under Microscopy," *Microscopy and Microanalysis*, vol. 5, pp. 528–50, 2004.

[3] A. Mattos, A. M. Sole-Cava, G. Decarli, and M. Benchimol, "Fine structure and isozymic characterization of trichomonadid protozoa," *Parasitology Research*, vol. 83, pp. 290–295, 1997.

[4] C. Yao and L. S. Köster, "Caprine prion gene polymorphisms are associated with decreased incidence of classical scrapie in goat herds in the United Kingdom," *Veterinary Research*, vol. 42, 2011.

[5] R. H. BonDurant, "Venereal Diseases of Cattle: Natural History, Diagnosis, and the Role of Vaccines in their Control," *Veterinary Clinics of North America: Food Animal Practice*, vol. 21, no. 2, pp. 383–408, 2005.

[6] K. C. Ribeiro, L. H. Monteiro-Leal, and M. Benchimol, "Contributions of the Axostyle and Flagella to Closed Mitosis in the Protists Tritrichomonas foetus and Trichomonas vaginalis," *The Journal of Eukaryotic Microbiology*, vol. 47, no. 5, pp. 481–492, 2000.

[7] M. Benchimol and F. Engelke, "Hydrogenosome behavior during the cell cycle in Tritrichomonas foetus," *Biology of the Cell*, vol. 95, no. 5, pp. 283–293, 2003.

[8] R. F. Madeiro Da Costa and M. Benchimol, "The effect of drugs on cell structure of Tritrichomonas foetus," *Parasitology Research*, vol. 92, no. 2, pp. 159–170, 2004.

[9] A. Pereira-Neves, C. M. Campero, A. Martínez, and M. Benchimol, "Identification of Tritrichomonas foetus pseudocysts in fresh preputial secretion samples from bulls," *Veterinary Parasitology*, vol. 175, no. 1-2, pp. 1–8, 2011.

[10] W. Amos, A. Grimstone, L. Rothschild, and R. Allen, "Structure, protein composition and birefringence of the costa: a motile flagellar root fibre in the flagellate Trichomonas," *Journal*

of Cell Science, vol. 35, no. 1, 1979.

[11] W. E. Sledge, A. D. Larson, and L. T. Hart, "Costae of Tritrichomonas foetus: Purification and chemical composition," *Science*, vol. 199, no. 4325, pp. 186–188, 1978.

[12] M. Rosa, I; de Souza, W; Benchimol, "High-resolution scanning electron microscopy of the cytoskeleton of Tritrichomonas foetus," *Journal of Structural Biology*, vol. 183, no. 3, pp. 412–418, 2013.

[13] Kstate, "Animal Parasitology https://www.k-state.edu/parasitology/625tutorials/Protozoa10.html (Accessed 19.5.20)."

[14] Cornell, "Tritrichomonas foetus: What shelters need to know – Gimme Shelter https://blogs.cornell.edu/cornellsheltermedicine/2014/07/01/tritrichomonas-foetus-illu/comment-page-1/ (Accessed 10.5.20)."

[15] P. Dolezal, O. Smid, P. Rada, Z. Zubácová, D. Bursać, R. Suták, J. Nebesárová, T. Lithgow, and J. Tachezy, "Giardia mitosomes and trichomonad hydrogenosomes share a common mode of protein targeting," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 31, pp. 10924–10929, 2005.

[16] J. Dąbrowska, J. Karamon, M. Kochanowski, J. Sroka, J. Zdybel, and T. Cencek, "Tritrichomonas foetus as a causative agent of tritrichomonosis in different animal hosts," *Journal of Veterinary Research*, vol. 63, no. 4, pp. 533–541, 2019.

[17] N. Yarlett and J. H. P. Hackstein, "Hydrogenosomes: One Organelle, Multiple Origins," *BioScience*, vol. 55, no. 8, pp. 657–668, 2005.

[18] F. Burki, "Mitochondrial Evolution: Going, Going, Gone," *Current Biology*, vol. 26, no. 10, pp. 410–12, 2016.

[19] G. A. Biagini, B. J. Finlay, and D. Lloyd, "Evolution of the hydrogenosome," *FEMS Microbiology Letters*, vol. 155, no. 2, pp. 133–140, 1997.

[20] R. P. Hirt, C. J. Noel, T. Sicheritz-Ponten, J. Tachezy, and P.-L. Fiori, "Trichomonas vaginalis surface proteins: a view from the genome.," *Trends in Parasitology*, vol. 23, no. 11, pp. 540–7, 2007.

[21] I. M. Parsonson, B. L. Clark, and J. H. Dufty, "Early pathogenesis and pathology of Tritrichomonas foetus infection in virgin heifers," *Journal of Comparative Pathology*, vol. 86, no. 1, pp. 59–66, 1976.

[22] J. Gaines, L. Paisley, and P. Anderson, "Trichomoniasis in a dairy herd: Control by artificial insemination and prostaglandin F2α treatment," *Theriogenology*, vol. 29, no. 6, pp. 1367–1374, 1988.

[23] Oklahoma-State-University, "Bovine Trichomoniasis https://extension.okstate.edu/fact-

sheets/bovine-trichomoniasis.html (Accessed 03.06.20)."

[24] E. R. Cobo, C. Morsella, D. Cano, A. Cipolla, and C. M. Campero, "Immunization in heifers with dual vaccines containing Tritrichomonas foetus and Campylobacter fetus antigens using systemic and mucosal routes," *Theriogenology*, vol. 62, no. 8, pp. 1367–1382, 2004.

[25] A. Samuelson, JD, Winter, "Bovine vibriosis: the nature of the carrier state in the bull. — Docphin," *Journal of Infectious Diseases*, vol. 116, no. 5, pp. 581–92, 1966.

[26] D. O. Rae and J. E. Crews, "Tritrichomonas foetus," *Veterinary Clinics of North America: Food Animal Practice*, vol. 22, no. 3, pp. 595–611, 2006.

[27] C. Yao, "Diagnosis of Tritrichomonas foetus-infected bulls, an ultimate approach to eradicate bovine trichomoniasis in US cattle?," *Journal of Medical Microbiology*, vol. 62, no. Pt_1, pp. 1–9, 2013.

[28] A. Yule, S. Z. Skirrowtand, and R. H. Bondurantf, "Bovine Trichomoniasis," *Parasitology Today*, vol. 5, no. I2, pp. 373–7, 1989.

[29] A. Pereira-Neves, L. Ferreira Nascimento, and M. Benchimol, "Cytotoxic Effects Exerted by Tritrichomonas foetus Pseudocysts," *Protist*, vol. 163, pp. 529–543, 2011.

[30] C. Yao and L. S. Köster, "Tritrichomonas foetus infection, a cause of chronic diarrhea in the domestic cat," *Veterinary Research*, vol. 46, no. 1, 2015.

[31] A. C. Rosypal, A. Ripley, H. D. Stockdale Walden, B. L. Blagburn, D. C. Grant, and D. S. Lindsay, "Survival of a feline isolate of Tritrichomonas foetus in water, cat urine, cat food and cat litter," *Veterinary Parasitology*, vol. 185, pp. 279–281, 2011.

[32] J. M. Maritz, K. M. Land, J. M. Carlton, and R. P. Hirt, "What is the importance of zoonotic trichomonads for human health?," *Trends in Parasitology*, vol. 30, no. 7, pp. 333–341, 2014.

[33] E. Mielczarek and J. Blaszkowska, "Trichomonas vaginalis: pathogenicity and potential role in human reproductive failure," *Infection*, vol. 44, no. 4, pp. 447–458, 2016.

[34] R. A. Martínez-Díaz, F. Ponce-Gordo, I. Rodríguez-Arce, M. C. del Martínez-Herrero, F. G. González, R. Á. Molina-López, and M. T. Gómez-Muñoz, "Trichomonas gypaetinii n. sp., a new trichomonad from the upper gastrointestinal tract of scavenging birds of prey," *Parasitology Research*, vol. 114, no. 1, pp. 101–112, 2015.

[35] G. A. Dennis Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "GenBank," *Nucleic Acids Research*, vol. 41, pp. 36–42, 2013.

[36] S. Kumar, G. Stecher, M. Li, C. Knyaz, and K. Tamura, "MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms," *Mol. Biol. Evol*, vol. 25, no. 1, 2018.

[37] S. Q. Le and O. Gascuel, "An Improved General Amino Acid Replacement Matrix," *Mol. Biol. Evol*, vol. 25, no. 7, 2008.

[38] R. S. Felleisen, "Host-parasite interaction in bovine infection with Tritrichomonas foetus," *Microbes and Infection*, vol. 1, no. 10, pp. 807–816, 1999.

[39] C. Profizi, A. Cian, D. Meloni, M. Hugonnard, V. Lambert, K. Groud, A.-C. Gagnon, E. Viscogliosi, and L. Zenner, "Prevalence of Tritrichomonas foetus infections in French catteries," *Veterinary Parasitology*, vol. 196, pp. 50–55, 2013.

[40] C. Yao, "Diagnosis of Tritrichomonas foetus-infected bulls, an ultimate approach to eradicate bovine trichomoniasis in US cattle?," *Journal of Medical Microbiology*, no. 62, 2013.

[41] Wedgewood, "Ronidazole for Veterinary Use https://www.wedgewoodpharmacy.com/learning-center/professional-monographs/ronidazole (Accessed 01/06/20)."

[42] V. Morin-Adeline, R. Lomas, D. O'Meally, C. Stack, A. Conesa, and J. Šlapeta, "Comparative transcriptomics reveals striking similarities between the bovine and feline isolates of Tritrichomonas foetus: consequences for in silico drug-target identification," *BMC Genomics*, vol. 15, no. 1, p. 955, 2014.

[43] H. S. Walden, C. Dykstra, A. Dillon, S. Rodning, D. Givens, R. Bird, J. Newton, and D. Lindsay, "A new species of Tritrichomonas (Sarcomastigophora: Trichomonida) from the domestic cat (Felis catus)," *Parasitology Research*, vol. 112, no. 6, pp. 2227–2235, 2013.

[44] Y. Jin, A. Du, and C. Yao, "Clinical isolates of Tritrichomonas foetus in bulls in Wyoming, South Dakota and Montana, USA," *BMC Veterinary Research*, vol. 16, no. 1, p. 12, 2020.

[45] J. Tachezy, R. Tachezy, V. Hampl, J. Flegr, and J. Kulda, "Cattle Pathogen Tritrichomonas foetus (Riedmuller, 1928) and Pig Commensal Tritrichomonas suis (Gruby & Delafond, 1843) Belong to the Same Species," *The Journal of Eukaryotic Microbiology*, vol. 49, no. 2, pp. 154–163, 2002.

[46] Z.-R. Lun, X.-G. Chen, X.-Q. Zhu, X.-R. Li, and M.-Q. Xie, "Are Tritrichomonas foetus and Tritrichomonas suis synonyms?," *Trends in Parasitology*, vol. 21, no. 3, pp. 122–125, 2005.

[47] C. P. Hibler, D. M. Hammond, F. H. Caskey, A. E. Johnson, and P. R. Fitzgerald, "The Morphology and Incidence of the Trichomonads of Swine, Tritrichomonas suis (Gruby and Delafond), Tritrichomonas rotunda, n. sp. And Trichomonas buttreyi, n.sp.," *The Journal of Protozoology*, vol. 7, no. 2, pp. 159–171, 1960.

[48] K. Mueller, V. Morin-Adeline, K. Gilchrist, G. Brown, and J. Slapeta, "High prevalence of Tritrichomonas foetus 'bovine genotype' in faecal samples from domestic pigs at a farm where bovine trichomonosis has not been reported for over 30 years," *Veterinary Parasitology*, vol. 212, pp. 105–110, 2015.

[49] J. Suzuki, S. Kobayashi, H. Osuka, D. Kawahata, T. Oishi, K. Sekiguchi, A. Hamada, and S. Iwata, "Characterization of a human isolate of Tritrichomonas foetus (cattle/swine geno-

type) infected by a zoonotic opportunistic infection," *Journal of Veterinary Internal Medicine*, vol. 78, no. 4, pp. 633–640, 2016.

[50] C. A. Zalonis, A. Pillay, W. Secor, B. Humburg, and R. Aber, "Rare case of trichomonal peritonitis," *Emerging Infectious Diseases*, vol. 17, no. 7, pp. 1312–3, 2011.

[51] C. F. Frey and N. Müller, "Tritrichomonas - Systematics of an enigmatic genus," *Molecular and Cellular Probes*, vol. 26, no. 3, pp. 132–136, 2012.

[52] T. Cavalier-Smith, "Kingdom Protozoa and Its 18 Phyla," Tech. Rep. 4, 1993.

[53] E. D. S. Wolff, S. W. Salisbury, J. R. Horner, and D. J. Varricchio, "Common Avian Infection Plagued the Tyrant Dinosaurs," *PLoS Genetics*, vol. 4, no. 9, 2009.

[54] M. Wenzel, R. Radek, G. Brugerolle, and H. Konig, "Indentification of the ectosymbiotic bacteria of Mixotricha paradoxa involved in movement symbiosis," *European Journal of Protisology*, vol. 29, pp. 11–23, 2003.

[55] R. S. J. Felleisen, "Comparative sequence analysis of 5n8S rRNA genes and internal transcribed spacer (ITS) regions of trichomonadid protozoa," *Parasitology*, vol. 115, pp. 111–119, 1997.

[56] A. S. Kucknoor, V. Mundodi, and J. F. Alderete, "Genetic identity and differential gene expression between Trichomonas vaginalis and Trichomonas tenax," *BMC Microbiology*, vol. 58, 2009.

[57] D. Gerbod, V. P. Edgcomb, C. Noel, Te, R. Wintjens, J. Tachezy, M. L. Sogin, and E. Viscogliosi, "Phylogenetic Relationships of Class II Fumarase Genes from Trichomonad Species," *Mol. Biol. Evol*, vol. 18, no. 8, pp. 1574–1584, 2001.

[58] D. Gerbod, E. Sanders, S. Moriya, and E. Viscogliosi, "Molecular phylogenies of Parabasalia inferred from four protein genes and comparison with rRNA trees," *Molecular Phylogenetics and Evolution*, vol. 31, pp. 572–580, 2004.

[59] E. Mirasol-Meléndez, L. G. Brieba, C. Díaz-Quezada, M. López-Hidalgo, E. E. Figueroa-Angulo, L. Ávila-González, R. Arroyo-Verástegui, and C. G. Benítez-Cardoza, "Characterization of multiple enolase genes from Trichomonas vaginalis. Potential novel targets for drug and vaccine design," *Parasitology International*, vol. 67, no. 4, pp. 444–453, 2018.

[60] Hirt RP and Sherrard J, "Trichomonas vaginalis origins, molecular pathobiology and clinical considerations," *Current Opinion in Infectious Diseases*, vol. 28, no. 1, 2015.

[61] J. R. Schwebke and D. Burgess, "Trichomoniasis.," *Clinical Microbiology Reviews*, vol. 17, no. 4, pp. 794–803, 2004.

[62] D. Soper, "Trichomoniasis: Under control or undercontrolled?," *American Journal of Obstetrics and Gynecology*, vol. 190, no. 1, pp. 281–290, 2004.

[63] N. Zhang, H. Zhang, Y. Yu, P. Gong, J. Li, Z. Li, T. Li, Z. Cong, C. Tian, X. Liu, X. Yu, and X. Zhang, "High prevalence of Pentatrichomonas hominis infection in gastrointestinal cancer patients," *Parasites Vectors*, vol. 12, p. 423, 2019.

[64] Z. F. Zhang, S. Graham, S. Z. Yu, J. Marshall, M. Zielezny, Y. X. Chen, M. Sun, S. L. Tang, C. S. Liao, J. L. Xu, and X. Z. Yang, "Trichomonas vaginalis and cervical cancer. A prospective study in China," *Annals of Epidemiology*, vol. 5, no. 4, pp. 325–332, 1995.

[65] J. R. Stark, G. Judson, J. F. Alderete, V. Mundodi, A. S. Kucknoor, E. L. Giovannucci, E. A. Platz, S. Sutcliffe, K. Fall, T. Kurth, J. Ma, M. J. Stampfer, and L. A. Mucci, "Prospective Study of Trichomonas vaginalis Infection and Prostate Cancer Incidence and Mortality: Physicians ' Health Study," *Journal of The National Cancer Institute*, vol. 101, no. 20, 2009.

[66] A. C. Fouts and S. J. Kraus, "Trichomonas vaginalis: Reevaluation of Its Clinical Presentation and Laboratory Diagnosis," Tech. Rep. 2, 1980.

[67] C. M. Ryan, N. De Miguel, and P. J. Johnson, "Trichomonas vaginalis: current understanding of host-parasite interactions," *Essays in Biochemistry*, vol. 51, pp. 161–175, 2011.

[68] E. Gould, L. Corbeil, S. Kania, and M. Tolbert, "Evaluation of surface antigen TF1.17 in feline Tritrichomonas foetus isolates," *Veterinary Parasitology*, vol. 244, pp. 144–153, 2017.

[69] L. C. Ribeiro, C. Santos, and M. Benchimol, "Is Trichomonas tenax a Parasite or a Commensal?," *Protist*, vol. 166, pp. 196–210, 2015.

[70] S. M. Hersh, "Pulmonary Trichomoniasis and Trichomonas tenax ," *Journal of Medical Microbiology*, vol. 20, pp. 1–10, 1985.

[71] H. Mallat, I. Podglajen, V. Lavarde, J.-L. Mainardi, J. Frappier, and M. Cornet, "Molecular Characterization of Trichomonas tenax Causing Pulmonary Infection," *Journal of Clinical Microbiology*, vol. 42, no. 8, pp. 3886–3887, 2004.

[72] W. W. Wantland and D. Lauer, "Correlation of Some Oral Hygiene Variables with Age, Sex, and Incidence of Oral Protozoa," *Journal of Dental Research*, vol. 49, no. 2, pp. 293–297, 1970.

[73] S. A. Ali Mohammed and A. B. Mohsen ALwaaly, "Prevalence trichomonas tenax in Karbala Governorate," *Journal of Physics Conference*, vol. 1294, 2019.

[74] P. Kleina, J. Bettim-Bandinelli, S. L. Bonatto, M. Benchimol, and M. R. Bogo, "Molecular phylogeny of Trichomonadidae family inferred from ITS-1, 5.8S rRNA and ITS-2 sequences," *International Journal for Parasitology*, vol. 34, no. 8, pp. 963–970, 2004.

[75] L. S. Garcia, "Dientamoeba fragilis, One of the Neglected Intestinal Protozoa," *Journal of Clinical Microbiology*, vol. 54, no. 9, pp. 2243–50, 2016.

[76] D. Stark, J. Barratt, D. Chan, and J. T. Ellis, "Dientamoeba fragilis, the Neglected Tri-

chomonad of the Human Bowel," *Clinical microbiology reviews*, vol. 29, no. 3, pp. 553–80, 2016.

[77] J. L. N. Barratt, J. Harkness, D. Marriott, J. T. Ellis, and D. Stark, "Gut Microbes A review of Dientamoeba fragilis carriage in humans: Several reasons why this organism should be considered in the diagnosis of gastrointestinal illness," *Gut Microbes*, vol. 3, pp. 3–12, 2011.

[78] J. L. Barratt, M. Cao, D. J. Stark, and J. T. Ellis, "The Transcriptome Sequence of Dientamoeba fragilis Offers New Biological Insights on its Metabolism, Kinome, Degradome and Potential Mechanisms of Pathogenicity," *Protist*, vol. 166, no. 4, pp. 389–408, 2015.

[79] E. M. Hussein, H. I. Al-Mohammed, and A. M. Hussein, "Genetic diversity of Dientamoeba fragilis isolates of irritable bowel syndrome patients by high-resolution melting-curve (HRM) analysis," *Parasitolgy Research*, vol. 105, no. 4, pp. 1053–60, 2009.

[80] D. Stark, J. Barratt, T. Roberts, D. Marriot, J. Harkness, and J. Ellis, " A Review of the Clinical Presentation of Dientamoebiasis," *Journal of Tropical Medicine and Hygiene*, vol. 84, no. 4, pp. 614–619, 2010.

[81] J. Intra, C. Sarto, S. Besana, N. Tiberti, and P. Brambilla, "The importance of considering the neglected intestinal protozoan parasite Dientamoeba fragilis," *Journal of Medical Microbiology*, vol. 68, pp. 890–892, 2019.

[82] L. Miguel, "Clinical and Epidemiological Characteristics of Patients with Dientamoeba fragilis Infection," *Am. J. Trop. Med. Hyg*, vol. 99, no. 5, pp. 1170–1173, 2018.

[83] A. Chudnovskiy, A. Mortha, V. Kana, Y. Belkaid, M. E. Grigg, and M. M. Correspondence, "Host-Protozoan Interactions Protect from Mucosal Infections through Activation of the Inflammasome," *Cell*, vol. 167, pp. 444–456, 2016.

[84] S. M. Cacciò, A. R. Sannella, E. Manuali, F. Tosini, M. Sensi, D. Crotti, and E. Pozio, "Pigs as natural hosts of Dientamoeba fragilis genotypes found in humans," *Emerging Infectious Diseases*, vol. 18, no. 5, pp. 838–841, 2012.

[85] D. Stark, N. Beebe, D. Marriott, J. Ellis, and J. Harkness, "Prospective Study of the Prevalence, Genotyping, and Clinical Relevance of Dientamoeba fragilis Infections in an Australian Population," *Journal of Clinical Microbiology*, vol. 43, no. 6, pp. 2718–2723, 2005.

[86] V. Munasinghe, N. Vella, J. Ellis, P. Windsor, and Stark. Damien, "Cyst formation and faecal–oral transmission of Dientamoeba fragilis– the missing link in the life cycle of an emerging pathogen," *International Journal for Parasitology*, vol. 43, pp. 879–883, 2013.

[87] D. Stark, L. S. Garcia, J. L. N. Barratt, O. Phillips, T. Roberts, D. Marriott, J. Harkness, and J. T. Ellis, "Description of Dientamoeba fragilis Cyst and Precystic Forms from Human Samples," *Journal of Clinical Microbiology*, vol. 52, no. 7, pp. 2680–2683, 2014.

[88] J. Slapeta, N. Müller, C. M. Stack, G. Walker, A. Lew-Tabor, J. Tachezy, and C. F. Frey, "Comparative analysis of Tritrichomonas foetus (Riedmüller, 1928) cat genotype, T. foetus (Riedmüller, 1928) cattle genotype and Tritrichomonas suis (Davaine, 1875) at 10 DNA loci," *International Journal for Parasitology*, vol. 42, no. 13-14, pp. 1143–1149, 2012.

[89] X. Li, J. Li, X. Zhang, Z. Yang, J. Yang, and P. Gong, "Prevalence of Pentatrichomonas hominis infections in six farmed wildlife species in Jilin, China," *Veterinary Parasitology*, vol. 244, pp. 160–163, 2017.

[90] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM," *arXiv*, 2013.

[91] W. Chao Li, K. Wang, and Y. Gu, "Occurrence of Blastocystis sp. and Pentatrichomonas hominis in sheep and goats in China," *Parasites Vectors*, vol. 11, no. 93, 2018.

[92] B. F. Bastos, B. Brener, M. Alves de Figueiredo, D. Leles, and F. Mendes-de Almeida, "Pentatrichomonas hominis infection in two domestic cats with chronic diarrhea," *Journal of Feline Medicine and Reports*, vol. 1, pp. 1–4, 2018.

[93] W. Zhang, G. Ren, W. Zhao, Z. Yang, Y. Shen, Y. Sun, and J. Cao, "Genotyping of Enterocytozoon bieneusi and Subtyping of Blastocystis in Cancer Patients: Relationship to Diarrhea and Assessment of Zoonotic Transmission," *Frontiers in Microbiology*, vol. 8, p. 1835, 2017.

[94] T. Tan, S. Ong, and K. Suresh, "Genetic variability of Blastocystis sp. isolates obtained from cancer and HIV/AIDS patients," *Parasitology Research*, vol. 105, pp. 1283–1286, 2009.

[95] C. Compaore, F. Lekpa, L. Nebie, P. Niamba, and A. Niakara, "Pentatrichomonas hominis infection in rheumatoid arthritis treated with adalimumab," *Rheumatology*, vol. 52, pp. 1534–1535, 2013.

[96] D. Meloni, C. Mantini, J. Goustille, and E. Viscogliosi, "Molecular identification of Pentatrichomonas hominis in two patients with gastrointestinal symptoms," *Journal of Clinical Pathology*, vol. 64, no. 10, pp. 933–933, 2011.

[97] S. Jongwutiwes, U. Silachamroon, and C. Putaporntip, "Pentatrichomonas hominis in empyema thoracis," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 94, no. 2, pp. 185–186, 2000.

[98] R. J. Lennon, J. C. Dunn, J. E. Stockdale, S. J. Goodman, A. J. Morris, and K. C. Hamer, "Trichomonad parasite infection in four species of Columbidae in the UK," *Parasitology*, vol. 140, no. 11, pp. 1368–1376, 2013.

[99] A. Amin, I. Bilic, D. Liebhart, and M. Hess, "Trichomonads in birds-a review," *Parasitology*, vol. 141, pp. 733–747, 2020.

[100] W. M. Harmon, W. A. Clark, A. C. Hawbecker, and M. Stafford, "Trichomonas gallinae in

columbiform birds from the Galapagos Islands.," *Journal of wildlife diseases*, vol. 23, no. 3, pp. 492–494, 1987.

[101] U. Hofle, C. Gortazar, J. A. Ortiz, B. Knispel, and E. F. Kaleta, "Outbreak of trichomoniasis in a woodpigeon (Columba palumbus) wintering roost," *European Journal of Wildlife Research*, vol. 50, no. 2, pp. 73–77, 2004.

[102] D. Villanúa, U. Höfle, L. Pérez-Rodríguez, and C. Gortázar, "Trichomonas gallinae in wintering Common Wood Pigeons Columba palumbus in Spain," *Ibis*, vol. 148, no. 4, pp. 641–648, 2006.

[103] J. M. Scimeca, D. E. Culberson, C. R. Abee, and W. A. Gardner, "Intestinal Trichomonads (Tritrichomonas mobilensis) in the Natural Host Saimiri sciureus and Saimiri boliviensis," *Veterinary Pathology*, vol. 26, pp. 144–147, 1989.

[104] M. D. C. Martínez-Herrero, M. M. Garijo-Toledo, F. González, I. Bilic, D. Liebhart, P. Ganas, M. Hess, and M. T. Gómez-Muñoz, "Membrane associated proteins of two Trichomonas gallinae clones vary with the virulence.," *PloS one*, vol. 14, no. 10, p. e0224032, 2019.

[105] R. A. Robinson, B. Lawson, M. P. Toms, K. M. Peck, and J. K. Kirkwood, "Emerging Infectious Disease Leads to Rapid Population Declines of Common British Birds," *PLoS ONE*, vol. 5, no. 8, p. 12215, 2010.

[106] R. M. Stabler, "Trichomonas gallinae: A review," *Experimental Parasitology*, vol. 3, no. 4, pp. 368–402, 1954.

[107] H. Mehlhorn, S. Al-Quraishy, A. Aziza, and M. Hess, "Fine structure of the bird parasites Trichomonas gallinae and Tetratrichomonas gallinarum from cultures," *Parasitology Research*, vol. 105, pp. 751–765, 2009.

[108] K. Bobrek, J. Urbanowicz, P. Chorbiński, and A. Gaweł, "Molecular analysis of Trichomonas gallinae in racing pigeons from Upper Silesia, Poland," *Polish Journal of Veterinary Sciences*, vol. 20, no. 1, pp. 185–187, 2017.

[109] J. Sansano-Maestre, M. Magdalena Garijo-Toledo, M. Teresa Gómez-Muñoz, and M. Teresa Gó mez-Muñ oz, "Avian Pathology Prevalence and genotyping of Trichomonas gallinae in pigeons and birds of prey Prevalence and genotyping of Trichomonas gallinae in pigeons and birds of prey," *Avian Pathology*, vol. 38, pp. 201–7, 2009.

[110] E. Grabensteiner, I. Bilic, T. Kolbe, and M. Hess, "Molecular analysis of clonal trichomonad isolates indicate the existence of heterogenic species present in different birds and within the same host," *Veterinary Parasitology*, vol. 172, no. 1, pp. 53–64, 2010.

[111] P. Babál and L. C. Russell, "Sialic Acid-Specific Lectin-Mediated Adhesion of Tritrichomonas foetus and," Tech. Rep. 1, 1999.

[112] P. Demes, F. F. Pindak, D. J. Wells, and W. A. Gardner, "Adherence and surface properties of Tritrichomonas mobilensis, an intestinal parasite of the squirrel monkey," *Parasitolgy Research*, vol. 75, pp. 589–594, 1989.

[113] V. Midlej, A. Pereira-Neves, L. W. Kist, M. R. Bogo, and M. Benchimol, "Ultrastructural features of Tritrichomonas mobilensis and comparison with Tritrichomonas foetus," *Veterinary Parasitology*, vol. 182, no. 2-4, pp. 171–180, 2011.

[114] Z. Zubacova, Z. Cim Urek, and J. Tachezy, "Molecular & Biochemical Parasitology Comparative analysis of trichomonad genome sizes and karyotypes," *Molecular & Biochemical Parasitology*, vol. 161, pp. 49–54, 2008.

[115] M. W. Lehker and J. F. Alderete, "Resolution of Six Chromosomes of Trichomonas vaginalis and Conservation of Size and Number among Isolates," *The Journal of Parasitology*, vol. 85, no. 5, p. 976, 1999.

[116] C. Gomez, M. E. Ramirez, M. Calixto-Galvez, O. Medel, and M. A. Rodríguez, "Regulation of Gene Expression in Protozoa Parasites," *Journal of Biomedicine and Biotechnology*, p. 24, 2010.

[117] S. A. Nadler and B. M. Honigberg, "Genetic Differentiation and Biochemical Polymorphism Among Trichomonads," *Journal of Parasitology*, vol. 74, no. 5, pp. 797–84, 1988.

[118] M. Tibayrenc, F. Kjellberg, and F. J. Ayala, "A clonal theory of parasitic protozoa: The population structures of Entamoeba, Giardia, Leishmania, Naegleria, Plasmodium, Trichomonas, and Trypanosoma and their medical and taxonomical consequences," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, pp. 2414–2418, 1990.

[119] W. D. Xu, Z. R. Lun, and A. Gajadhar, "Chromosome numbers of Tritrichomonas foetus and Tritrichomonas suis," *Veterinary Parasitology*, vol. 78, no. 4, pp. 247–251, 1998.

[120] J. Dabrowska, I. Keller, J. Karamon, M. Kochanowski, B. Gottstein, T. Cencek, C. F. Frey, and N. Müller, "Whole genome sequencing of a feline strain of Tritrichomonas foetus reveals massive genetic differences to bovine and porcine isolates," *International Journal for Parasitology*, vol. 20, 2020.

[121] Bertram, "A Comparison of Isozymes of Five Axenic Giardia Isolates - PubMed," *Journal of Parasitology*, vol. 69, no. 5, pp. 793–81, 1983.

[122] T. E. Nash, T. Mccutchan, D. Keister, J. B. Dame, J. D. Conrad, and F. D. Gillin, "Restriction-Endonuclease Analysis of DNA from 15 Giardia Isolates Obtained from Humans and Animals," *The Journal of Infectious Diseases*, vol. 152, no. 1, 1985.

[123] M. D. Conrad, M. Bradic, S. D. Warring, A. W. Gorman, and J. M. Carlton, "Getting trichy: tools and approaches to interrogating Trichomonas vaginalis in a post-genome world," *Trends*

in Parasitology, vol. 29, 2012.

[124] Y.-K. Fang, K.-Y. Chien, K.-Y. Huang, W.-H. Cheng, F.-M. Ku, R. Lin, T.-W. Chen, P.-J. Huang, C.-H. Chiu, and P. Tang, "Responding to a Zoonotic Emergency with Multi-omics Research: Pentatrichomonas hominis Hydrogenosomal Protein Characterization with Use of RNA Sequencing and Proteomics," *OMICS: A Journal of Integrative Biology*, vol. 20, no. 11, pp. 662–669, 2016.

[125] J. M. Carlton, R. P. Hirt, J. C. Silva, A. L. Delcher, M. Schatz, Q. Zhao, J. R. Wortman, S. L. Bidwell, U. Cecilia, M. Alsmark, S. Besteiro, T. Sicheritz-Ponten, C. J. Noel, J. B. Dacks, P. G. Foster, C. Simillion, Y. Van De Peer, D. Miranda-Saavedra, G. J. Barton, G. D. Westrop, S. Müller, D. Dessi, P. L. Fiori, Q. Ren, I. Paulsen, H. Zhang, F. D. Bastida, M. Embley, G. H. Coombs, J. C. Mottram, J. Tachezy, C. M. Fraser-Liggett, P. J. Johnson, and J. Craig, "Draft Genome Sequence of the Sexually Transmitted Pathogen Trichomonas vaginalis," *Science*, vol. 11, no. 3155809, pp. 207–212, 2007.

[126] R. E. Schneider, M. T. Brown, A. M. Shiflett, S. D. Dyall, R. D. Hayes, Y. Xie, J. A. Loo, and P. J. Johnson, "The Trichomonas vaginalis hydrogenosome proteome is highly reduced relative to mitochondria, yet complex compared with mitosomes," *International Journal for Parasitology*, vol. 41, no. 13-14, pp. 1421–1434, 2011.

[127] Missouri, "Reproductive Anatomy and Physiology of the Cow — MU Extension https://extension2.missouri.edu/g2015 (Accessed 2.4.20)."

[128] U. Beef, "Pregnant cows, timing of pregnancy, open cows, pregnancy rate — UNL Beef."

[129] Beef-Magazine, "What To Do With Open Cows — Beef Magazine https://www.beefmagazine.com/cow-calf/what-do-open-cows (Accessed 4.3.20)."

[130] Dairexnet, "Pregnant vs. Open: Getting Cows Pregnant and the Money it Makes – https://dairy-cattle.extension.org/pregnant-vs-open-getting-cows-pregnant-and-the-money-it-makes/ (Acessed 22.05.20)."

[131] Y. T. Grohn and P. J. Rajala-Schultz, "Epidemiology of reproductive performance in dairy cows," in *Animal Reproduction Science* (Y. Grohn and P. Rajala-Schultz, eds.), vol. 60-61, pp. 605–614, Elsevier, 2000.

[132] T. B. Ault, B. A. Clemmons, S. T. Reese, F. G. Dantas, G. A. Franco, T. P. L. Smith, J. Lannett Edwards, P. R. Myer, and K. G. Pohler, "Bacterial taxonomic composition of the postpartum cow uterus and vagina prior to artificial insemination 1," *Journal of Animal Science*, vol. 97, no. 3, pp. 4305–4313, 2019.

[133] S. Witkin and I. Linhares, "Why do lactobacilli dominate the human vaginal microbiota?," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 124, no. 4, pp. 606–611,

2017.

[134] D. J. Miller, "Review: The epic journey of sperm through the female reproductive tract," *Animal*, vol. 12, pp. 27–39, 2018.

[135] I. M. Sheldon, J. G. Cronin, G. D. Healey, C. Gabler, W. Heuwieser, D. Streyl, J. J. Bromfield, A. Miyamoto, C. Fergani, and H. Dobson, "Innate immunity and inflammation of the bovine female reproductive tract in health and disease," *Reproduction*, vol. 148, pp. 41–51, 2014.

[136] M. J. Yaeger and J. L. Gookin, "Histologic features associated with Tritrichomonas foetus-induced colitis in domestic cats," *Veterinary Pathology*, vol. 42, no. 6, pp. 797–804, 2005.

[137] J. D. Ondrak, "Tactics for Identifying and Eliminating Tritrichomonas foetus from Tactics for Identifying and Eliminating Tritrichomonas foetus from Infected Beef Herds Infected Beef Herds," tech. rep., 2010.

[138] D. Rae, J. E. Crews, E. C. Greiner, and G. Donovan, "Epidemiology of Tritrichomonas foetus in beef bull populations in Florida," *Theriogenology*, vol. 61, no. 4, pp. 605–618, 2004.

[139] L. Ball, D. A. Dargatz, J. M. Cheney, and R. G. Mortimer, "Control of Venereal Disease in Infected Herds," *Veterinary Clinics of North America: Food Animal Practice*, vol. 3, pp. 561–574, 1987.

[140] B. N. Singh, J. J. Lucas, G. R. Hayes, I. Kumar, D. H. Beach, M. Frajblat, R. O. Gilbert, U. Sommer, and C. E. Costello, "Tritrichomonas foetus Induces Apoptotic Cell Death in Bovine Vaginal Epithelial Cells," *Infection and Immunity*, vol. 72, no. 7, pp. 4151–4158, 2004.

[141] Y. Shi, W. Jiang, Z. Ma, and Y. Qiu, "A case report of pulmonary tritrichomonosis in a pig," *BMC Veterinary Research*, vol. 13, 2017.

[142] J. A. Mendoza-Ibarra, S. Pedraza-Díaz, F. J. García-Peña, S. Rojo-Montejo, J. A. Ruiz-Santa-Quiteria, E. San Miguel-Ibáñez, V. Navarro-Lozano, L. M. Ortega-Mora, K. Osoro, and E. Collantes-Fernandez, "High prevalence of Tritrichomonas foetus infection in Asturiana de la Montaña beef cattle kept in extensive conditions in Northern Spain," *The Veterinary Journal*, vol. 193, no. 1, pp. 146–151, 2012.

[143] L. Molina, J. Perea, G. Meglia, E. Angón, and A. García, "Spatial and temporal epidemiology of bovine trichomoniasis and bovine genital campylobacteriosis in La Pampa province (Argentina)," *Preventive Veterinary Medicine*, vol. 110, no. 3-4, pp. 388–394, 2013.

[144] B. Clark, D. Emery, and J. Dufty, "Therapeutic immunisation of bulls with the membranes and glycoproteins of Trifrichomonas foetus var. brisbane," *Australian Veterinary Journal*, vol. 61, no. 2, pp. 65–66, 1984.

[145] W. Kvasnicka, R. Taylor, J.-c. Huang, R. Tronstad, A. Bosomworth, and M. Hall, "The-

riogenology Investigations of the Incidence of bovine Trichomoniasis in Nevada and of the Efficacy of Immunizing Cattle with Vaccines Containing Trichomonas foetus.," *Theriogenology*, vol. 31, no. 5, pp. 963–971, 1989.

[146] Animal-Science, "Texas has the largest cow herds in the US - Animal Science https://animalscience.tamu.edu/2019/04/12/texas-has-the-largest-cow-herds-in-the-u-s/ (Accessed 03.06.20)."

[147] W. Beef, "World Beef Consumption Per Capita (Ranking of Countries) - Beef2Live — Eat Beef * Live Better https://beef2live.com/story-world-beef-consumption-per-capita-ranking-countries-0-111634 (Accessed 20.5.20)."

[148] Statista, "Dairy industry in Europe - Statistics & Facts — Statista https://www.statista.com/topics/3955/dairy-industry-in-europe/ (Accessed 20.5.20)."

[149] OIE, "OIE - World Organisation for Animal Health https://www.oie.int/standard-setting/terrestrial-manual/access-online/ (Accessed 27.5.20)."

[150] M. A. Edmondson, K. S. Joiner, J. A. Spencer, K. P. Riddell, S. P. Rodning, J. A. Gard, and M. D. Givens, "Impact of a killed Tritrichomonas foetus vaccine on clearance of the organism and subsequent fertility of heifers following experimental inoculation," *Theriogenology*, vol. 90, pp. 245–251, 2017.

[151] F. Dufernez, R. L. Walker, C. Noel, C. Stephanie, and M. Clea, "Morphological and Molecular Identification of Non-Tritrichomonas foetus Trichomonad Protozoa from the Bovine Preputial Cavity," *The Journal of Eukaryotic Microbiology*, vol. 54, no. 2, pp. 161–168, 2007.

[152] P. C. Irons, M. M. Henton, and H. J. Bertschinger, "Collection of preputial material by scraping and aspiration for the diagnosis of Tritrichomonas foetus in bulls," *Journal of the South African Veterinary Association*, vol. 73, no. 2, pp. 66–69, 2002.

[153] J. C. Rhyan, P. C. Blanchard, W. G. Kvasnicka, M. R. Hall, and D. Hanks, "Tissue-invasive Tritrichomonas foetus in four aborted bovine fetuses," tech. rep., 1995.

[154] Colorado State, "Parasitology http://csu-cvmbs.colostate.edu/vdl/parasitology/Pages/default.aspx (Accessed 09.06.20)."

[155] R. Bhatt, M. Abraham, and G. E. Garber, "New concepts in the diagnosis and pathogenesis of Trichomonas vaginalis," *The Canadian Journal of Infectious Diseases*, vol. 7, no. 5, 1996.

[156] E. Mielczarek and J. Blaszkowska, "Trichomonas vaginalis: pathogenicity and potential role in human reproductive failure," *Infection*, vol. 44, no. 4, pp. 447–58, 2016.

[157] A. Warton and B. Honigberg, "Lectin Analysis of Surface Saccharides in Two Trichomonas vaginalis Strains Differing in Pathogenicity," *The Journal of Protozoology*, vol. 27, no. 4, pp. 410–419, 1980.

[158] V. B. Kon, J. M. Papadimitriou, T. A. Robertson, and A. Warton, "Quantitation of concanavalin A and wheat germ agglutinin binding by two strains of Trichononas vaginalis of differing pathogenicity using gold particle-conjugated lectins," *Parasitology Research*, vol. 75, no. 1, pp. 7–13, 1988.

[159] F. D. Bastida-Corcuera, C. Y. Okumura, A. Colocoussi, and P. J. Johnson, "Trichomonas vaginalis lipophosphoglycan mutants have reduced adherence and cytotoxicity to human ectocervical cells," *Eukaryotic Cell*, vol. 4, no. 11, pp. 1951–1958, 2005.

[160] R. N. Fichorova, R. T. Trifonova, R. O. Gilbert, C. E. Costello, G. R. Hayes, J. J. Lucas, and B. N. Singh, "Trichomonas vaginalis lipophosphoglycan triggers a selective upregulation of cytokines by human female reproductive tract epithelial cells," *Infection and Immunity*, vol. 74, no. 10, pp. 5773–5779, 2006.

[161] N. De Miguel, G. Lustig, O. Twu, A. Chattopadhyay, J. A. Wohlschlegel, and P. J. Johnson, "Proteome Analysis of the Surface of Trichomonas vaginalis Reveals Novel Proteins and Strain-dependent Differential Expression," *Molecular & Cellular Proteomics*, vol. 9, no. 7, pp. 1554–66, 2010.

[162] A. F. Garcia and J. F. Alderete, "Characterization of the Trichomonas vaginalis surface-associated AP65 and binding domain interacting with trichomonads and host cells," *BMC Microbiology*, vol. 7, p. 116, 2007.

[163] J. A. Engbring and J. F. Alderete, "Three genes encode distinct AP33 proteins involved in Trichomonas vaginalis cytoadherence," *Molecular Microbiology*, vol. 28, no. 2, pp. 305–313, 1998.

[164] K. Kim, "Role of proteases in host cell invasion by Toxoplasma gondii and other Apicomplexa," *Acta Tropica*, vol. 91, no. 1, pp. 69–81, 2004.

[165] C. J. Noël, N. Diaz, T. Sicheritz-Ponten, L. Safarikova, J. Tachezy, P. Tang, P.-L. Fiori, and R. P. Hirt, "Trichomonas vaginalis vast BspA-like gene family: evidence for functional diversity from structural organisation and transcriptomics," *BMC Genomics*, vol. 8, no. 11, 2010.

[166] P. H. Davis, Z. Zhang, M. Chen, X. Zhang, S. Chakraborty, and S. L. Stanley, "Identification of a family of BspA like surface proteins of Entamoeba histolytica with novel leucine rich repeats," *Molecular and Biochemical Parasitology*, vol. 145, no. 1, pp. 111–116, 2006.

[167] M. R. Handrich, S. G. Garg, E. W. Sommerville, R. P. Hirt, and S. B. Gould, "BspA and Pmp proteins of Trichomonas vaginalis mediate adherence to host cells," *bioRxiv*, 2018.

[168] G. A. De Carli, P. Brasseur, A. C. Da Silva, A. Wendorff, and M. Rott, "Hemolytic Activity of Trichomonas vaginalis and Tritrichomonas foetus," *Memorias do Instituto Oswaldo Cruz*,

vol. 91, no. 1, pp. 107–110, 1996.

[169] P. L. Fiori, P. Rappelli, A. M. Rocchigiani, and P. Cappuccinelli, "Trichomonas vaginalis haemolysis: Evidence of functional pores formation on red cell membranes," *FEMS Microbiology Letters*, vol. 109, pp. 378–1097, 1993.

[170] J. B. De Jesus, P. Cuervo, C. Britto, L. Sabóia-Vahia, F. C. E. Suva-Filho, A. Borges-Veloso, D. B. Petrópolis, E. Cupolillo, and G. B. Domont, "Cysteine peptidase expression in trichomonas vaginalis isolates D isplaying High- And low-virulence phenotypes," *Journal of Proteome Research*, vol. 8, no. 3, pp. 1555–1564, 2009.

[171] P. Cuervo, E. Cupolillo, C. Britto, L. J. González, F. C. e Silva-Filho, L. C. Lopes, G. B. Domont, and J. B. De Jesus, "Differential soluble protein expression between Trichomonas vaginalis isolates exhibiting low and high virulence phenotypes," *Journal of Proteomics*, vol. 71, no. 1, pp. 109–122, 2008.

[172] Y. R. Nievas, V. M. Coceres, ·. V. Midlej, ·. Wanderley De Souza, M. Benchimol, A. Pereira-Neves, A. A. Vashisht, J. A. Wohlschlegel, P. J. Johnson, and N. De Miguel, "Membrane-shed vesicles from the parasite Trichomonas vaginalis: characterization and their association with cell interaction," *Cellular and Molecular Life Sciences*, vol. 75, pp. 2211–2226, 2018.

[173] D. He, G. Pengtao, Y. Ju, L. Jianhua, L. He, Z. Guocai, and Z. Xichen, "Differential protein expressions in virus-infected and uninfected trichomonas vaginalis," *Korean Journal of Parasitology*, vol. 55, no. 2, pp. 121–128, 2017.

[174] V. Margarita, A. Marongiu, N. Diaz, D. Dessì, P. L. Fiori, and P. Rappelli, "Prevalence of double-stranded RNA virus in Trichomonas vaginalis isolated in Italy and association with the symbiont Mycoplasma hominis," *Parasitology Research*, vol. 118, no. 12, pp. 3565–3570, 2019.

[175] K. J. Graves, A. P. Ghosh, N. Schmidt, P. Augostini, W. E. Secor, J. R. Schwebke, D. H. Martin, P. J. Kissinger, and C. A. Muzny, "Trichomonas vaginalis Virus Among Women With Trichomoniasis and Associations With Demographics, Clinical Outcomes, and Metronidazole Resistance," *Clinical Infectious Diseases ®*, vol. 35233, no. 12, pp. 2170–2176, 2018.

[176] K. J. Graves, A. P. Ghosh, P. J. Kissinger, and C. A. Muzny, "Trichomonas vaginalis virus: a review of the literature," *International Journal of STD and AIDS*, vol. 30, no. 5, pp. 496–504, 2019.

[177] E. R. Cobo, L. B. Corbeil, and R. H. BonDurant, "Immunity to infections in the lower genital tract of bulls," *Journal of Reproductive Immunology*, vol. 89, no. 1, pp. 55–61, 2011.

[178] L. Corbeil, "Vaccination strategies against Tritrichomonas foetus," *Parasitology Today*, vol. 10, no. 3, pp. 103–106, 1994.

[179] M. K. Tolbert and J. L. Gookin, "Mechanisms of Tritrichomonas foetus Pathogenicity in Cats with Insights from Venereal Trichomonosis," *Journal of Veterinary Internal Medicine*, vol. 30, no. 2, pp. 516–26, 2016.

[180] J. C. Rhyan, K. L. Wilson, B. Wagner, M. L. Anderson, R. H. Bondurant, D. E. Burgess, G. K. Mutwiri, and L. B. Corbeil, "Demonstration of Tritrichomonas foetus in the External Genitalia and of Specific Antibodies in Preputial Secretions of Naturally Infected Bulls," *Veterinary Pathology*, vol. 36, pp. 406–411, 1999.

[181] L. B. Corbeil, M. L. Anderson, R. R. Corbeil, J. M. Eddow, and R. H. BonDurant, "Female Reproductive Tract Immunity in Bovine Trichomoniasis," *American Journal of Reproductive Immunology*, vol. 39, no. 3, pp. 189–198, 1998.

[182] P. Soto and A. E. Parma, "The Immune Response in Cattle Infected with Tri tric ho mo nas Foetus," *Veterinary Parasitology*, vol. 33, pp. 343–348, 1989.

[183] J. S. Ikeda, R. H. Bondurant, C. M. Campero, and L. B. Corbeill, "Conservation of a Protective Surface Antigen of Tritrichomonas foetus," *Journal of Clinical Microbiology*, pp. 3289–3295, 1993.

[184] J. Gookin, S. Stauffer, D. Dybas, and D. Cannon, "Documentation of In Vivo and In Vitro Aerobic Resistance of Feline Tritrichomonas foetus Isolates to Ronidazole," *Journal of Veterinary Internal Medicine*, vol. 24, no. 4, pp. 1003–1007, 2010.

[185] Bi-vetmedica https://www.bi-vetmedica.com/species/cattle/products/TrichGuard.html (Accessed 09.06.20), "TrichGuard® — Boehringer Ingelheim Vetmedica."

[186] Janzen, "Overview of Trichomoniasis in Cattle - Reproductive System - Veterinary Manual https://www.msdvetmanual.com/reproductive-system/trichomoniasis/overview-of-trichomoniasis-in-cattle (Accessed 03/06/20)."

[187] L. I. Fuchs, M. C. Fort, D. Cano, C. M. Bonetti, H. D. Giménez, P. M. Vázquez, D. Bacigalupe, J. D. Breccia, C. M. Campero, and J. A. Oyhenart, "Clearance of Tritrichomonas foetus in experimentally infected heifers protected with vaccines based on killed-T. foetus with different adjuvants," *Vaccine*, vol. 35, no. 9, pp. 1341–1346, 2017.

[188] E. R. Cobo, D. Cano, O. Rossetti, and C. M. Campero, "Heifers immunized with whole-cell and membrane vaccines against Tritrichomonas foetus and naturally challenged with an infected bull," *Veterinary Parasitology*, vol. 109, no. 3-4, pp. 169–184, 2002.

[189] W. G. Kvasnicka, D. Hanks, J. C. Huang, M. R. Hall, D. Sandblom, H. J. Chu, L. Chavez, and W. M. Acree, "Clinical evaluation of the efficacy of inoculating cattle with a vaccine containing Tritrichomonas foetus," *American Journal of Veterinary Research*, vol. 53, no. 11, pp. 2023–2027, 1992.

[190] C. Ribeiro, M. Falleiros, S. Bicudo, J. Araujo Júnior, M. Golim, F. Silva Filho, C. Padovani, and J. Modolo, "Tritrichomonas fetus extracellular products decrease progressive motility of bull sperm," *Theriogenology*, vol. 73, no. 1, pp. 64–70, 2010.

[191] C. I. Shaia, J. Voyich, S. J. Gillis, B. N. Singh, and D. E. Burgess, "Purification and Expression of the Tf190 Adhesin in Tritrichomonas foetus," *Infection and Immunity*, vol. 66, no. 3, pp. 1100–1105, 1998.

[192] S. Herr, L. M. M. Ribeiro, E. Claassen, and J. G. Myburgh, "A Reduction in the Duration of Infection with Tritrichomonas Foetus Following Vaccination in Heifers and the Failure to Demonstrate a Curative Effect in Infected Bulls," *Journal of Veterinary Research*, vol. 58, pp. 41–45, 1991.

[193] P. Baltzell, H. Newton, and A. O'Connor, "A Critical Review and Meta-Analysis of the Efficacy of Whole-Cell Killed Tritrichomonas foetus Vaccines in Beef Cattle," *Journal of Veterinary Internal Medicine*, vol. 27, no. 4, pp. 760–770, 2013.

[194] Y. Xie, J. Gong, M. Li, H. Fang, and W. Xu, "The Medicinal Potential of Influenza Virus Surface Proteins: Hemagglutinin and Neuraminidase," *Current Medicinal Chemistry*, vol. 18, no. 7, pp. 1050–1066, 2011.

[195] M. J. Sylte and D. L. Suarez, "Influenza Neuraminidase as a Vaccine Antigen," in *Vaccines for Pandemic Influenza*, pp. 227–241, Springer, Berlin, Heidelberg, 2009.

[196] S. A. Plotkin, "Vaccines, Vaccination, and Vaccinology," *The Journal of Infectious Diseases*, vol. 187, no. 9, pp. 1349–1359, 2003.

[197] R. Rappuoli, M. J. Bottomley, U. D'Oro, O. Finco, and E. De Gregorio, "Reverse vaccinology 2.0: Human immunology instructs vaccine antigen design," *The Journal of Experimental Medicine*, vol. 213, no. 4, pp. 469–481, 2016.

[198] R. Moxon, P. A. Reche, and R. Rappuoli, "Reverse Vaccinology," *Frontiers in Immunology*, vol. 10, 2019.

[199] M. J. Rodríguez-Ortega, N. Norais, G. Bensi, S. Liberatori, S. Capo, M. Mora, M. Scarselli, F. Doro, G. Ferrari, I. Garaguso, T. Maggi, A. Neumann, A. Covre, J. L. Telford, and G. Grandi, "Characterization and identification of vaccine candidate proteins through analysis of the group A Streptococcus surface proteome," *Nature Biotechnology*, vol. 24, no. 2, pp. 191–197, 2006.

[200] A. Sette and R. Rappuoli, "Reverse vaccinology: developing vaccines in the era of genomics.," *Immunity*, vol. 33, no. 4, pp. 530–41, 2010.

[201] V. Masignani, M. Pizza, and E. R. Moxon, "The development of a vaccine against Meningococcus B using reverse vaccinology," *Frontiers in Immunology*, vol. 10, no. MAR, p. 751,

2019.

[202] I. Bianconi, B. Alcalá-Franco, M. Scarselli, M. Dalsass, S. Buccato, A. Colaprico, S. Marchi, V. Masignani, and A. Bragonzi, "Genome-Based Approach Delivers Vaccine Candidates Against Pseudomonas aeruginosa," *Frontiers in Immunology*, vol. 9, p. 3021, 2019.

[203] P. Olin, F. Rasmussen, L. Gustafsson, H. O. Hallander, and H. Heijbel, "Randomised controlled trial of two-component, three-component, and five-component acellular pertussis vaccines compared with whole-cell pertussis vaccine," *Lancet*, vol. 350, no. 9091, pp. 1569–1577, 1997.

[204] A. Naz, F. M. Awan, A. Obaid, S. A. Muhammad, R. Z. Paracha, J. Ahmad, and A. Ali, "Identification of putative vaccine candidates against Helicobacter pylori exploiting exoproteome and secretome: A reverse vaccinology based approach," *Infection, Genetics and Evolution*, vol. 32, pp. 280–291, 2015.

[205] K. L. Seib, X. Zhao, and R. Rappuoli, "Developing vaccines in the era of genomics: a decade of reverse vaccinology," *Clin Microbiol Infect*, vol. 18, pp. 109–116, 2012.

[206] A. V. S. Hill, "Vaccines against malaria," *Philosophical Transactions of The Royal Society*, no. 366, 2011.

[207] P. E. Duffy, T. Sahu, A. Akue, N. Milman, and C. Anderson, "Pre-erythrocytic malaria vaccines: identifying the targets," *Expert Rev Vaccines*, vol. 11, no. 10, pp. 1261–1280, 2012.

[208] K. Couper, K. L. Flanagan, A. Frimpong, W. Ndifon, K. A. Kusi, and M. Fokuo Ofori, "Novel Strategies for Malaria Vaccine Design THE GLOBAL MALARIA SITUATION," *Frontiers in Immunology — www.frontiersin.org*, vol. 9, p. 2769, 2018.

[209] S. Sundar and B. Singh, "Identifying vaccine targets for anti-leishmanial vaccine development," *Expert Rev Vaccines*, vol. 13, no. 4, 2014.

[210] E. Del Tordello, R. Rappuoli, and I. Delany, "Reverse Vaccinology," in *Human Vaccines*, pp. 65–86, Elsevier, 2017.

[211] H. Etz, D. B. Minh, T. Henics, A. Dryla, B. Winkler, C. Triska, A. P. Boyd, J. Söllner, W. Schmidt, U. Von Ahsen, M. Buschle, S. R. Gill, J. Kolonay, H. Khalak, C. M. Fraser, A. Von Gabain, E. Nagy, and A. Meinke, "Identification of in vivo expressed vaccine candidate antigens from Staphylococcus aureus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 10, pp. 6573–6578, 2002.

[212] D. P. Knox, D. L. Redmond, G. F. Newlands, P. J. Skuce, D. Pettit, and W. D. Smith, "The nature and prospects for gut membrane proteins as vaccine candidates for Haemonchus contortus and other ruminant trichostrongyloids," *International Journal for Parasitology*, vol. 33, no. 11, pp. 1129–1137, 2003.

[213] G. Grandi, "Bacterial surface proteins and vaccines Introduction and context," *Biology Reports*, vol. 2, 2010.

[214] A. Olaya-Abril, I. Jiménez-Munguía, L. Gómez-Gascón, I. Obando, and M. J. Rodríguez-Ortega, "Identification of Potential New Protein Vaccine Candidates through Pan-Surfomic Analysis of Pneumococcal Clinical Isolates from Adults," *PLoS ONE*, vol. 8, no. 7, 2013.

[215] M. H. Chiang, W. C. Sung, S. P. Lien, Y. Z. Chen, A. F. yun Lo, J. H. Huang, S. C. Kuo, and P. Chong, "Identification of novel vaccine candidates against Acinetobacter baumannii using reverse vaccinology," *Human Vaccines and Immunotherapeutics*, vol. 11, no. 4, pp. 1065–1073, 2015.

[216] E. H. Nardin and R. S. Nussenzweig, "Stage-specific antigens on the surface membrane of sporozoites of malaria parasites," *Nature*, vol. 274, no. 5666, pp. 55–57, 1978.

[217] R. Strugnell, F. Zepp, A. Cunningham, and T. Tantawichien, "Vaccine antigens," *Perspectives in Vaccinology*, vol. 1, no. 1, pp. 61–88, 2011.

[218] D. Chargelegue, P. M. Drake, P. Obregon, A. Prada, N. Fairweather, and J. K. Ma, "Highly immunogenic and protective recombinant vaccine candidate expressed in transgenic plants," *Infection and Immunity*, vol. 73, no. 9, pp. 5915–5922, 2005.

[219] Á. D. L. C. Pech-Canul, V. Monteón, and R. L. Solís-Oviedo, "A Brief View of the Surface Membrane Proteins from Trypanosoma cruzi," *Journal of Parasitology Research*, vol. 1, 2017.

[220] X. Ge, T. Kitten, C. L. Munro, D. H. Conrad, and P. Xu, "Pooled Protein Immunization for Identification of Cell Surface Antigens in Streptococcus sanguinis," *PLoS ONE*, vol. 5, no. 7, 2010.

[221] A. Emms, "OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences," *bioRxiv*, 2018.

[222] D. M. Emms and S. Kelly, "OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy," *Genome Biology*, no. 157, 2015.

[223] Pacific Biosciences, "SMRT Analysis Software - PacBio https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/ (Accessed 22.02.2018)."

[224] A. Rhoads and K. F. Au, "PacBio Sequencing and Its Applications," *Genomics, Proteomics & Bioinformatics*, vol. 13, pp. 278–289, 2015.

[225] R. M. Waterhouse, M. Seppey, F. A. Sim˜ Ao, M. Manni, P. Ioannidis, G. Klioutchnikov, E. V. Kriventseva, and E. M. Zdobnov, "BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics," *Molecular Biology and Evolution*, vol. 3, pp. 543–548, 2018.

[226] F. A. Sima, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs," *Bioinformatics*, vol. 19, pp. 3210–2, 2015.

[227] M. Benchimol, L. G. P. De Almeida, A. T. Vasconcelos, I. De Andrade Rosa, M. R. Bogo, E. Luiza, W. Kist, and W. De Souza, "Draft Genome Sequence of Tritrichomonas foetus Strain K," *Genome Announcements*, vol. 5, no. 16, 2017.

[228] M. Bradic, S. D. Warring, G. E. Tooley, P. Scheid, W. E. Secor, K. M. Land, P.-J. Huang, T.-W. Chen, C.-C. Lee, P. Tang, S. A. Sullivan, and J. M. Carlton, "Genetic Indicators of Drug Resistance in the Highly Repetitive Genome of Trichomonas vaginalis.," *Genome biology and evolution*, vol. 9, no. 6, pp. 1658–1672, 2017.

[229] M. M. Waldrop, "More Than Moore," *Nature*, vol. 530, pp. 144–147, 2016.

[230] S. J. Goodswen, P. J. Kennedy, and J. T. Ellis, "Evaluating High-Throughput Ab Initio Gene Finders to Discover Proteins Encoded in Eukaryotic Pathogen Genomes Missed by Laboratory Techniques," *PLoS ONE*, vol. 7, no. 11, 2012.

[231] M. Stanke and B. Morgenstern, "AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints," *Nucleic Acids Research*, vol. 1, no. 33, 2005.

[232] A. D. Johnson, R. E. Handsaker, S. L. Pulit, M. M. Nizzari, C. J. O 'donnell, P. I. W. De Bakker, and A. Bateman, "SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap," *Bioinformatics Applications*, vol. 24, no. 24, pp. 2938–293910, 2008.

[233] K. J. Hoff, S. Lange, A. Lomsadze, M. Borodovsky, and M. Stanke, "BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1.," *Bioinformatics*, vol. 32, no. 5, pp. 767–769, 2016.

[234] J. Besemer, A. Lomsadze, and M. Borodovsky, "GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions," Tech. Rep. 12, 2001.

[235] K. Hoff, A. Lomsadze, M. Borodovsky, and M. Stanke, "Whole-Genome Annotation with BRAKER," *Methods in Molecular Biology*, vol. 1962, pp. 65–95, 2019.

[236] I. De Andrade Rosa, G. Atella, and M. Benchimol, "Tritrichomonas foetus Displays Classical Detergent-resistant Membrane Microdomains on its Cell Surface," *Annals of Anatomy*, vol. 165, pp. 293–304, 2014.

[237] Y.-C. Liao, S.-H. Lin, and H.-H. Lin, "Completing bacterial genome assemblies: strategy and performance comparisons," *Scientific Reports*, vol. 5, 2015.

[238] L. G. P. Almeida, R. Paixão, R. C. Souza, G. C. Da Costa, F. J. A. Barrientos, M. Trindade Dos Santos, D. F. De Almeida, A. Tereza, and R. Vasconcelos, "A System for Automated Bac-

terial (genome) Integrated Annotation-SABIA," *Bioinformatics Applications Note*, vol. 20, no. 16, pp. 2832–2833, 2004.

[239] M. Bradic, S. D. Warring, G. E. Tooley, P. Scheid, W. E. Secor, K. M. Land, P.-J. Huang, T.-W. Chen, C.-C. Lee, P. Tang, S. A. Sullivan, and J. M. Carlton, "Genetic Indicators of Drug Resistance in the Highly Repetitive Genome of Trichomonas vaginalis," *Genome Biology and Evolution*, vol. 9, no. 6, pp. 1658–72, 2017.

[240] D. R. Zerbino and E. Birney, "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs," *Genome research*, vol. 18, no. 5, pp. 821–9, 2008.

[241] D. R. Zerbino, "Using the Velvet de novo assembler for short-read sequencing technologies," *Curr Protoc Bioinformatics*, 2010.

[242] G. D. Westrop, L. Wang, G. J. Blackburn, T. Zhang, L. Zheng, D. G. Watson, and G. H. Coombs, "Metabolomic profiling and stable isotope labelling of Trichomonas vaginalis and Tritrichomonas foetus reveal major differences in amino acid metabolism including the production of 2-hydroxyisocaproic acid, cystathionine and S-methylcysteine," *PLoS ONE*, vol. 3, 2017.

[243] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, no. 4, pp. 357–359, 2012.

[244] J. Kong, S. Huh, J. I. Won, J. Yoon, B. Kim, and K. Kim, "GAAP: A Genome Assembly + Annotation Pipeline," *BioMed Research International*, 2019.

[245] M. Chorev and L. Carmel, "The function of introns," *Frontiers in Genetics*, vol. 3, no. APR, p. 55, 2012.

[246] B.-S. Jo and S. S. Choi, "Introns: The Functional Benefits of Introns in Genomes," *Genomics & Informatics*, vol. 13, no. 4, p. 112, 2015.

[247] A. M. McGuire, M. D. Pearson, D. E. Neafsey, and J. E. Galagan, "Cross-kingdom patterns of alternative splicing and splice recognition," *Genome Biology*, vol. 9, no. 3, 2008.

[248] C. B. Menezes, A. P. Frasson, and T. Tasca, "Trichomoniasis – are we giving the deserved attention to the most common non-viral sexually transmitted disease worldwide?," *Microbial Cell*, vol. 3, no. 9, pp. 404–419, 2016.

[249] F. Mercer and P. J. Johnson, "Trichomonas vaginalis: Pathogenesis, Symbiont Interactions, and Host Cell Immune Responses," *Trends in Parasitology*, vol. 34, no. 8, pp. 683–693, 2018.

[250] F. D. Bastida-Corcuera, B. N. Singh, G. C. Gray, P. D. Stamper, M. Davuluri, K. Schlangen, R. R. Corbeil, and L. B. Corbeil, "Antibodies to Trichomonas vaginalis surface glycolipid," *Sexually Transmitted Infections*, vol. 89, no. 6, pp. 467–472, 2013.

[251] Y. P. Chen, A. M. Riestra, A. K. Rai, and P. J. Johnson, "A novel cadherin-like protein

mediates adherence to and killing of host cells by the parasite trichomonas vaginalis," *mBio*, vol. 10, no. 3, 2019.

[252] L. Horvathova, L. S. Afarikova, M. Basler, I. Hrdy, N. B. Campo, J.-W. Shin, K.-Y. Huang, P.-J. Huang, R. Lin, P. Tang, and J. Tachezy, "Transcriptomic Identification of Iron-Regulated and Iron-Independent Gene Copies within the Heavily Duplicated Trichomonas vaginalis Genome," *Genome Biol Evol*, vol. 4, no. 10, pp. 1017–1029, 2012.

[253] M. M. Mostegl, B. Richter, N. Nedorost, C. Lang, A. Maderner, N. Dinhopl, and H. Weissenböck, "First evidence of previously undescribed trichomonad species in the intestine of pigs?," *Veterinary Parasitology*, vol. 185, no. 2-4, pp. 86–90, 2012.

[254] V. Hampl, A. Pavlicek, and J. Flegr, "Construction and bootstrap analysis of DNA fingerprinting-based phylogenetic trees with a freeware program FreeTree: Application to trichomonad parasites," *International Journal of Systematic and Evolutionary Microbiology*, no. 51, pp. 731–5, 2001.

[255] A. Fahad Alrefaei, R. Low, N. Hall, R. Jardim, A. D. Avila, R. Gerhold, S. John, S. Steinbiss, A. A. Cunningham, B. Lawson, D. Bell, and K. Tyler, "Multilocus Analysis Resolves the European Finch Epidemic Strain of Trichomonas gallinae and Suggests Introgression from Divergent Trichomonads," *Genome Biology and Evolution*, vol. 11, no. 8, pp. 2391–2402, 2019.

[256] N. T. Reem, H. Y. Chen, M. Hur, X. Zhao, E. S. Wurtele, X. Li, L. Li, and O. Zabotina, "Comprehensive transcriptome analyses correlated with untargeted metabolome reveal differentially expressed pathways in response to cell wall alterations," *Plant Molecular Biology*, vol. 96, no. 4-5, pp. 509–529, 2018.

[257] H. Tan, J. Zhang, X. Qi, X. Shi, J. Zhou, X. Wang, and X. Xiang, "Correlation analysis of the transcriptome and metabolome reveals the regulatory network for lipid synthesis in developing Brassica napus embryos," *Plant Molecular Biology*, vol. 99, no. 1-2, pp. 31–44, 2019.

[258] M. C. Frith, M. Pheasant, and J. S. Mattick, "The amazing complexity of the human transcriptome," *European Journal of Human Genetics*, vol. 13, no. 8, pp. 894–897, 2005.

[259] S. K. Kushwaha, V. Kesarwani, S. Choudhury, S. Gandhi, and S. Sharma, "SARS-CoV-2 transcriptome analysis and molecular cataloguing of immunodominant epitopes for multi-epitope based vaccine design," *bioRxiv*, p. 2020.05.14.097170, 2020.

[260] A. Zawawi, R. Forman, H. Smith, I. Mair, M. Jibril, M. H. Albaqshi, A. Brass, J. P. Derrick, and K. J. Else, "In silico design of a T-cell epitope vaccine candidate for parasitic helminth infection," *PLOS Pathogens*, vol. 16, no. 3, 2020.

[261] J. Tuju, G. Kamuyu, L. M. Murungi, and F. H. A. Osier, "Vaccine candidate discovery for the next generation of malaria vaccines," *Immunology*, vol. 152, pp. 195–206, 2017.

[262] R. Furtado, M. Da Costa, and M. Benchimol, "The effect of drugs on cell structure of Tritrichomonas foetus," *Parasitology Today*, no. 92, pp. 159–170, 2004.

[263] C. Castro, R. Figueiredo, S. Menna-Barreto, N. De, S. Fernandes, L. Saboia-Vahia, G. Dias-Lopes, C. Britto, P. Cuervo, J. Batista, and D. Jesus, "Iron-modulated pseudocyst formation in Tritrichomonas foetus," *Parasitology*, vol. 143, no. 8, 2017.

[264] H. J. Atkinson, J. H. Morris, T. E. Ferrin, and P. C. Babbitt, "Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies," *PLoS ONE*, vol. 4, no. 2, p. 4345, 2009.

[265] T. Frickey and A. Lupas, "CLANS: a Java application for visualizing protein families based on pairwise similarity," *Bioinformatics Applications Note*, vol. 20, no. 18, pp. 3702–3704, 2004.

[266] S. Cheng, S. Karkar, E. Bapteste, N. Yee, P. Falkowski, and D. Bhattacharya, "Sequence similarity network reveals the imprints of major diversification events in the evolution of microbial life," *Frontiers in Ecology and Evolution*, vol. 2, p. 72, 2014.

[267] D. Beaudet, Y. Terrat, S. Bastien Halary, I. Enrique De La Providencia, and M. Hijri, "Mitochondrial Genome Rearrangements in Glomus Species Triggered by Homologous Recombination between Distinct mtDNA Haplotypes," *Genome Biology and Evolution*, vol. 5, no. 9, 2013.

[268] Hardy-Diagnostics, "Modified Diamonds Medium - for Trichomonas https://catalog.hardydiagnostics.com/cp_prod/Content/hugo/ModDiamondsMed.htm (Accessed 10.6.20)."

[269] Pacific Biosciences, "HGAP in SMRT Analysis http://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP-in-SMRT-Analysis (Accessed 22.02.2018)," 2017.

[270] Pacific Biosciences, "De Novo Assembly - Pacific Biosciences http://www.pacb.com/products-and-services/analytical-software/smrt-analysis/analysis-applications/de-novo-assembly/ (Accessed 03.08.2017)," 2017.

[271] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy, "Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation," *Genome Research*, vol. 27, no. 5, pp. 722–736, 2017.

[272] S. Steinbiss, F. Silva-Franco, B. Brunk, B. Foth, C. Hertz-Fowler, M. Berriman, and T. D. Otto, "Companion: a web server for annotation and analysis of parasite genomes," *Nucleic Acids Research*, vol. 44, 2016.

[273] T. D. Otto, G. P. Dillon, W. S. Degrave, and M. Berriman, "RATT: Rapid Annotation Transfer Tool," *Nucleic Acids Research*, vol. 39, no. 9, 2011.

[274] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg, "Improved microbial gene identification with GLIMMER," Tech. Rep. 23, 1999.

[275] A. L. Delcher, K. A. Bratke, E. C. Powers, and S. L. Salzberg, "Identifying bacterial genes and endosymbiont DNA with Glimmer," *Bioinformatics*, vol. 23, no. 6, pp. 673–679, 2007.

[276] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M.-A. Ele Rajandream, and B. Barrell, "Artemis: sequence visualization and annotation," *Bioinformatics Applications*, vol. 16, no. 10, pp. 944–945, 2000.

[277] Splice, "RNA info: Splice site consensus https://science.umd.edu/labs/mount/RNAinfo/consensus.html (Accessed 10.6.20)."

[278] NCBI, "RefSeq non-redundant proteins https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/ (Accessed 10.6.20)."

[279] S. Gö Tz, J. M. García-Gó Mez, J. Terol, T. D. Williams, S. H. Nagaraj, M. J. Nueda, M. Robles, M. Taló N, J. Dopazo, and A. Conesa, "High-throughput functional annotation and data mining with the Blast2GO suite," *Nucleic Acids Research*, vol. 36, no. 10, pp. 3420–3435, 2008.

[280] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles, "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research," *Bioinformatics Applications Note*, vol. 21, no. 18, pp. 3674–3676, 2005.

[281] P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez, and S. Hunter, "InterProScan 5: genome-scale protein function classification.," *Bioinformatics (Oxford, England)*, vol. 30, no. 9, pp. 1236–40, 2014.

[282] R. D. Finn, T. K. Attwood, P. C. Babbitt, A. Bateman, P. Bork, A. J. Bridge, H.-Y. Chang, Z. Dosztányi, S. El-Gebali, M. Fraser, J. Gough, D. Haft, G. L. Holliday, H. Huang, X. Huang, I. Letunic, R. Lopez, S. Lu, A. Marchler-Bauer, H. Mi, J. Mistry, D. A. Natale, M. Necci, G. Nuka, C. A. Orengo, Y. Park, S. Pesseat, D. Piovesan, S. C. Potter, N. D. Rawlings, N. Redaschi, L. Richardson, C. Rivoire, A. Sangrador-Vegas, C. Sigrist, I. Sillitoe, B. Smithers, S. Squizzato, G. Sutton, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, I. Xenarios, L.-S. Yeh, S.-Y. Young, and A. L. Mitchell, "InterPro in 2017-beyond protein family and domain annotations.," *Nucleic Acids Research*, vol. 45, no. D1, pp. 190–199, 2017.

[283] C. J. A Sigrist, E. de Castro, L. Cerutti, B. A. atrice Cuche, N. Hulo, A. Bridge, L. Bougueleret, and I. Xenarios, "New and continuing developments at PROSITE," *Nucleic acids re-*

search, vol. 41, pp. 344–7, 2013.

[284] P. D. Thomas, A. Kejariwal, M. J. Campbell, H. Mi, K. Diemer, N. Guo, I. Ladunga, B. Ulitsky-Lazareva, A. Muruganujan, S. Rabkin, J. A. Vandergriff, and O. Doremieux, "PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification.," *Nucleic acids research*, vol. 31, no. 1, pp. 334–41, 2003.

[285] M. Necci, D. Piovesan, Z. Dosztányi, S. C. Tosatto, and A. Valencia, "MobiDB-lite: Fast and highly specific consensus prediction of intrinsic disorder in proteins," *Bioinformatics*, vol. 33, no. 9, 2017.

[286] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[287] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "KEGG as a reference resource for gene and protein annotation," *Nucleic Acids Research*, vol. 44, pp. 457–62, 2016.

[288] M. Kanehisa, Y. Sato, and K. Morishima, "BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences," *Journal of Molecular Biology*, vol. 428, no. 2, pp. 726–731, 2016.

[289] E. L. L. Sonnhammer, G. Von Heijne, and A. Krogh, "A hidden Markov model for predicting transmembrane helices in protein sequences," *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, vol. 6, pp. 175–82, 1998.

[290] A. Krogh, B. È. Rn Larsson, G. Von Heijne, and E. L. L. Sonnhammer, "Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes," *Journal of Molecular Biology*, vol. 305, pp. 567–580, 2001.

[291] L. Kall, A. Krogh, and E. L. Sonnhammer, "A combined transmembrane topology and signal peptide prediction method," *Journal of Molecular Biology*, vol. 338, no. 5, pp. 1027–1036, 2004.

[292] O. Emanuelsson, S. Brunak, G. Von Heijne, and H. Nielsen, "Locating proteins in the cell using TargetP, SignalP and related tools," *Nature Protocols*, vol. 2, no. 4, pp. 953–971, 2007.

[293] A. Pierleoni, P. L. Martelli, and R. Casadio, "PredGPI: a GPI-anchor predictor," *BMC bioinformatics*, vol. 9, no. 392, 2008.

[294] T. Kinoshita, "Glycosylphosphatidylinositol (GPI) anchors: Biochemistry and cell biology: Introduction to a thematic review series," *Journal of Lipid Research*, vol. 57, no. 1, pp. 4–5, 2016.

[295] B. Beckwith-Cohen, O. Koren, S. Blum, , and D. Elad, "Variations in Vaginal pH in Dairy Cattle Associated with Parity and the Periparturient Period," *Israel Journal of Veterinary*

Medicine, vol. 67, no. 1, 2012.

[296] D. Lloyd, "The Effects of Oxygen on Fermentation in Tritrichomonas foetus KVl and its Variant 1MR-100 with Defective Hydrogenosomes," *Journal of General Microbiology*, vol. 133, no. 1, pp. 181–1, 1987.

[297] S. R. Mack and M. Müller, "Effect of Oxygen and Carbon Dioxide on the Growth of Trichomonas vaginalis and Tritrichomonas foetus," *Journal of Parasitology*, vol. 64, no. 5, pp. 927–929, 1978.

[298] Y. Wang, B. N. Ametaj, D. J. Ambrose, and M. G. Gänzle, "Characterisation of the bacterial microbiota of the vagina of dairy cows and isolation of pediocin-producing Pediococcus acidilactici," *BMC Microbiology*, vol. 13, no. 19, 2013.

[299] N. Rohland and D. Reich, "Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture," *Genome Research*, vol. 22, no. 5, pp. 939–946, 2012.

[300] Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads https://journal.embnet.org/index.php/embnetjournal/article/view/200/458 (Accessed 10.06.20)."

[301] N. Joshi and J. Fass, "GitHub - sickle: Windowed Adaptive Trimming for fastq files using quality https://github.com/najoshi/sickle (Accessed 10.06.20)."

[302] E. Afgan, D. Baker, M. van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, B. Grüning, A. Guerler, J. Hillman-Jackson, G. Von Kuster, E. Rasche, N. Soranzo, N. Turaga, J. Taylor, A. Nekrutenko, and J. Goecks, "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update," *Nucleic Acids Research*, vol. 44, no. 1, pp. 3–10, 2016.

[303] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype," *Nature Biotechnology*, vol. 37, no. 8, pp. 907–915, 2019.

[304] D. Kim, B. Langmead, and S. L. Salzberg, "HISAT: A fast spliced aligner with low memory requirements," *Nature Methods*, vol. 12, no. 4, pp. 357–360, 2015.

[305] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup, "The Sequence Alignment/Map format and SAMtools.," *Bioinformatics*, vol. 25, no. 16, pp. 2078–9, 2009.

[306] W. Bioinformatics, "WEHI Bioinformatics - featureCounts http://bioinf.wehi.edu.au/featureCounts/ (Accessed 3.8.2017)."

[307] RStudio, "RStudio – Open source and enterprise https://www.rstudio.com/ (Accessed 03.08.2017)," 2012.

[308] R-Project, "R: The R Project for Statistical Computing https://www.r-project.org/ (Accessed 03.08.2017)," 2017.

[309] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, 2010.

[310] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A software Environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[311] P. P. Chan and T. M. Lowe, "tRNAscan-SE: Searching for tRNA genes in genomic sequences," in *Methods in Molecular Biology*, vol. 1962, pp. 1–14, Humana Press Inc., 2019.

[312] T. M. Lowe and P. P. Chan, "tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes," *Web Server issue Published online*, vol. 44, 2016.

[313] K. Lagesen, P. Hallin, A. Rødland, H.-H. Staerfeldt, T. Rognes, and D. W. Ussery, "RNAmmer: consistent and rapid annotation of ribosomal RNA genes," *Nucleic Acids Research*, vol. 35, no. 9, pp. 3100–3108, 2007.

[314] C. Woehle, G. Kusdian, C. Radine, D. Graur, G. Landan, and S. B. Gould, "The parasite Trichomonas vaginalis expresses thousands of pseudogenes and long non-coding RNAs independently from functional neighbouring genes," *BMC Genomics*, vol. 15, 2014.

[315] S. Suzuki, M. Kakuta, T. Ishida, and Y. Akiyama, "GHOSTX: An Improved Sequence Homology Search Algorithm Using a Query Suffix Array and a Database Suffix Array," *PLoS ONE*, vol. 9, no. 8, 2014.

[316] J. Barratt, R. Gough, D. Stark, and J. Ellis, "Bulky Trichomonad Genomes: Encoding a Swiss Army Knife," 2016.

[317] Y.-K. Fang, K.-Y. Huang, P.-J. Huang, R. Lin, M. Chao, and P. Tang, "Gene-expression analysis of cold-stress response in the sexually transmitted protist Trichomonas vaginalis," *Journal of Microbiology*, vol. 48, no. 6, pp. 662–75, 2015.

[318] K.-Y. Huang, Y.-Y. Margaret Chen, Y.-K. Fang, W.-H. Cheng, C.-C. Cheng, Y.-C. Chen, T. E. Wu, F.-M. Ku, S.-C. Chen, R. Lin, and P. Tang, "Adaptive responses to glucose restriction enhance cell survival, antioxidant capability, and autophagy of the protozoan parasite Trichomonas vaginalis," *Biochimia Biophysica Acta*, vol. 1840, no. 1, pp. 53–64, 2014.

[319] A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton, "Jalview Version 2—a multiple sequence alignment editor and analysis workbench," *Bioinformatics Applications*, vol. 25, no. 910, pp. 1189–1191, 2009.

[320] L. K. Johnson, H. Alexander, and C. Titus Brown, "Re-assembly, quality evaluation, and

annotation of 678 microbial eukaryotic reference transcriptomes," *GigaScience*, vol. 8, pp. 1–12, 2018.

[321] T. Nordahl Petersen, S. Brunak, G. von Heijne, and H. Nielsen, "SignalP 4.0: discriminating signal peptides from transmembrane regions," *Nature Publishing Group*, vol. 8, pp. 785–786, 2011.

[322] H. Nielsen, "Predicting Secretory Proteins with SignalP," *Protein Function Prediction*, pp. 59–73, 2017.

[323] H. J. Atkinson, P. C. Babbitt, and M. Sajid, "The global cysteine peptidase landscape in parasites," *Trends in Parasitology*, vol. 25, no. 12, pp. 573–581, 2009.

[324] J. Von Dwingelo, I. Y. W. Chung, C. T. Price, L. Li, S. Jones, M. Cygler, and Y. A. Kwaik, "Interaction of the Ankyrin H core effector of legionella with the host LARP7 component of the 7SK snRNP complex," *mBio*, vol. 10, no. 4, 2019.

[325] G. H. Coombs, G. D. Westrop, P. Suchan, G. Puzova, R. P. Hirt, T. Martin Embley, J. C. Mottram, and S. Mü ller, "The Amitochondriate Eukaryote Trichomonas vaginalis Contains a Divergent Thioredoxin-linked Peroxiredoxin Antioxidant System," *Journal of Biological Chemistry*, vol. 279, no. 7, pp. 5249–56, 2004.

[326] K. Judge, M. Hunt, S. Reuter, A. Tracey, M. A. Quail, J. Parkhill, and S. J. Peacock, "Comparison of bacterial genome assembly software for MinION data and their applicability to medical microbiology," *Microbial Genomics*, vol. 2, no. 9, 2016.

[327] M. R. Handrich, S. G. Garg, E. W. Sommerville, R. P. Hirt, and S. B. Gould, "Characterization of the BspA and Pmp protein family of trichomonads," *Parasites Vectors*, vol. 12, p. 406, 2019.

[328] Y. Zhen and H. Stenmark, "Cellular functions of Rab GTPases at a glance," *Journal of Cell Science*, vol. 128, no. 17, pp. 3171–3176, 2015.

[329] D. E. Bosch and D. P. Siderovski, "Entamoeba histolytica RacC selectively engages p21-activated kinase effectors," *Biochemistry*, vol. 54, no. 2, pp. 404–412, 2015.

[330] V. Morin-Adeline, K. Mueller, A. Conesa, and J. Slapeta, "Comparative RNA-seq analysis of the Tritrichomonas foetus PIG30/1 isolate from pigs reveals close association with Tritrichomonas foetus BP-4 isolate 'bovine genotype'," *Veterinary Parasitology*, vol. 212, pp. 111–117, 2015.

[331] S. L. Salzberg, "Next-generation genome annotation: We still struggle to get it right," 2019.

[332] S. E. Brenner, "Errors in genome annotation," *Trends in Genetics*, vol. 15, no. 4, pp. 132–133, 1999.

[333] Lia and E. M. De Jesus, Ribeiro, "Human Immune Responses to Schistosoma mansoni Vaccine

Candidate Antigens," *Infection and Immunity*, vol. 68, no. 5, pp. 2797–2803, 2000.

[334] N. Khalaf Alharbi, I. Qasim, A. Almasoud, H. A. Aljami, M. W. Alenazi, A. Alhafufi, O. S. Aldibasi, A. M. Hashem, S. Kasem, R. Albrahim, M. Aldubaib, A. Almansour, N. J. Temperton, A. Kupke, S. Becker, A. Abu-obaidah, A. Alkarar, I.-K. Yoon, E. Azhar, T. Lambe, F. Bayoumi, A. Aldowerij, O. H. Ibrahim, S. C. Gilbert, and H. H. Balkhy, "Humoral immunogenicity and Efficacy of a Single Dose of ChAdOx1 MERS Vaccine Candidate in Dromedary Camels," *Scientific Reports*, vol. 9, pp. 1–11, 2019.

[335] B. Aslam, M. Basit, M. Atif Nisar, M. Khurshid, and M. H. Rasool, "Proteomics: Technologies and Their Applications," *Journal of Chromatographic Science*, vol. 55, no. 2, pp. 182–196, 2017.

[336] S. A. Hayes, S. Clarke, N. Pavlakis, and V. M. Howell, "The role of proteomics in the age of immunotherapies," *Mammalian Genome*, vol. 29, no. 11, pp. 757–769, 2018.

[337] K. Novakova, O. Sedo, and Z. Zdrahal, "Mass Spectrometry Characterization of Plant Phosphoproteins," *Current Protein & Peptide Science*, vol. 12, no. 2, pp. 112–125, 2011.

[338] S. Sundar and B. Singh, "Understanding Leishmania parasites through proteomics and implications for the clinic," *Expert Review Proteomics*, vol. 15, no. 5, pp. 371–390, 2018.

[339] P. J. Myler and K. D. Stuart, "Recent developments from the Leishmania genome project," *Current Opinion in Microbiology*, vol. 3, no. 4, pp. 412–416, 2000.

[340] B. Ockenfels, E. Michael, and M. A. Mcdowell, "Meta-analysis of the Effects of Insect Vector Saliva on Host Immune Responses and Infection of Vector-Transmitted Pathogens: A Focus on Leishmaniasis," *PLoS Negl Trop Dis*, vol. 8, no. 10, p. 3197, 2014.

[341] P.-R. Karhemo, S. Ravela, M. Laakso, I. Ritamo, O. Tatti, S. Mäkinen, S. Goodison, U.-H. Stenman, E. Hölttä, S. Hautaniemi, L. Valmu, K. Lehti, and P. Laakkonen, "An optimized isolation of biotinylated cell surface proteins reveals novel players in cancer metastasis," *Journal of Proteomics*, vol. 21, pp. 87–100, 2012.

[342] Rabilloud Thierry, "Membrane proteins and proteomics: Love is possible, but so difficult," *Electrophoresis*, vol. 30, pp. 1074–80, 2009.

[343] N. A. Haverland, M. Waas, I. Ntai, T. Keppel, R. L. Gundry, and N. L. Kelleher, "Cell Surface Proteomics of N-Linked Glycoproteins for Typing of Human Lymphocytes," *Proteomics*, vol. 17, no. 19, 2017.

[344] M. M. Shimogawa, E. A. Saada, A. A. Vashisht, W. D. Barshop, J. A. Wohlschlegel, K. L. Hill, and K. L. H. Analyzed Data, "Cell Surface Proteomics Provides Insight into Stage-Specific Remodeling of the Host-Parasite Interface in Trypanosoma brucei," *Molecular & Cellular Proteomics*, vol. 14, pp. 1977–1988, 1977.

[345] B. J. Davids, C. M. Liu, E. M. Hanson, C. H. Y. Le, J. Ang, K. Hanevik, M. Fischer, M. Radunovic, N. Langeland, M. Ferella, S. G. Svärd, M. Ghassemian, Y. Miyamoto, and L. Eckmann, "Identification of Conserved Candidate Vaccine Antigens in the Surface Proteome of Giardia lamblia," *Infection and Immunity*, vol. 21, no. 87, 2019.

[346] L. J. Stroud, J. Slapeta, M. P. Padula, D. Druery, G. Tsiotsioras, J. R. Coorssen, and C. M. Stack, "Comparative proteomic analysis of two pathogenic Tritrichomonas foetus genotypes: there is more to the proteome than meets the eye," *International Journal for Parasitology*, vol. 47, pp. 203–213, 2017.

[347] D. Chelius and P. V. Bondarenko, "Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry," *Journal of Proteome Research*, vol. 1, no. 4, pp. 317–323, 2002.

[348] L. N. Darville and B. H. Sokolowski, "Label-free quantitative mass spectrometry analysis of differential protein expression in the developing cochlear sensory epithelium," *Proteome Science*, vol. 16, no. 1, 2018.

[349] V. J. Patel, K. Thalassinos, S. E. Slade, J. B. Connolly, A. Crombie, J. C. Murrell, and J. H. Scrivens, "A Comparison of Labeling and Label-Free Mass Spectrometry-Based Proteomics Approaches," *Journal of Proteome Research*, vol. 8, pp. 3752–3759, 2009.

[350] J. Guo and L. Prokai, "To tag or not to tag: A comparative evaluation of immunoaffinity-labeling and tandem mass spectrometry for the identification and localization of posttranslational protein carbonylation by 4-hydroxy-2-nonenal, an end-product of lipid peroxidation," *Journal of Proteomics*, vol. 74, no. 11, pp. 2360–2369, 2011.

[351] S.-E. Ong and M. Mann, "Stable Isotope Labeling by Amino Acids in Cell Culture for Quantitative Proteomics," in *Quantitative Proteomics by Mass Spectrometry*, pp. 37–52, Humana Press, 2007.

[352] K. Sap and J. Demmers, "Labeling Methods in Mass Spectrometry Based Quantitative Proteomics," in *Integrative Proteomics*, InTech, 2012.

[353] L. Dayon and J. C. Sanchez, "Relative protein quantification by MS/MS using the tandem mass tag technology," *Methods in Molecular Biology*, vol. 893, pp. 115–127, 2012.

[354] National Institutes of Health https://ods.od.nih.gov/factsheets/Biotin-HealthProfessional/ (08.07.20), "Biotin - Health Professional Fact Sheet," 2020.

[355] G. F. Combs, J. P. McClung, G. F. Combs, and J. P. McClung, "Chapter 15 – Biotin," in *The Vitamins*, pp. 371–385, Academic Press, 2017.

[356] GBiosciences https://info.gbiosciences.com/blog/the-advantages-of-biotinylation-tagging-in-protein-purification (Accessed 08.07.20), "The Advantages of Biotinylation Tagging in Protein

Purification," 2018.

[357] Molecular Diagnostic Services http://www.mds-usa.com/avidbiomethods.html (Accessed 08.07.21), "Avidin Biotin Methods," 2011.

[358] C. Niehage, J. Karbanová, C. Steenblock, D. Corbeil, and B. Hoflack, "Cell Surface Proteome of Dental Pulp Stem Cells Identified by Label-Free Mass Spectrometry," *PloS one*, vol. 11, no. 8, 2016.

[359] Y. Luo, K. Mcdonald, and J. W. Hanrahan, "Trafficking of immature F508-CFTR to the plasma membrane and its detection by biotinylation," *Biochem. J*, vol. 419, pp. 211–219, 2009.

[360] C. Niehage, C. Steenblock, T. Pursche, M. Bornhä User, D. Corbeil, and B. Hoflack, "The Cell Surface Proteome of Human Mesenchymal Stromal Cells," *PloS one*, vol. 6, no. 5, 2011.

[361] L. J. Foster, P. A. Zeemann, C. Li, M. Mann, O. N. Jensen, and M. Kassem, "Differential Expression Profiling of Membrane Proteins by Quantitative Proteomics in a Human Mesenchymal Stem Cell Line Undergoing Osteoblast Differentiation," *Stem Cells*, vol. 23, no. 9, pp. 1367–1377, 2005.

[362] H. Gong, K. Kobayashi, T. Sugi, H. Takemae, T. Horimoto, X. Xuan, H. Akashi, and K. Kato, "Pull-down method to access the cell surface receptor for Toxoplasma gondii," *Parasitology International*, vol. 65, no. 5, pp. 514–515, 2016.

[363] J. Masuoka, L. N. Guthrie, and K. C. Hazen, "Complications in cell-surface labelling by biotinylation of Candida albicans due to avidin conjugate binding to cell-wall proteins," *Microbiology*, vol. 148, no. 4, pp. 1073–1079, 2002.

[364] P. V. Pham, "Medical biotechnology: Techniques and applications," in *Omics Technologies and Bio-engineering*, vol. 1, pp. 449–469, Elsevier Inc., 2018.

[365] C.-J. Hu, G. Song, W. Huang, G.-Z. Liu, C.-W. Deng, H.-P. Zeng, L. Wang, F.-C. Zhang, X. Zhang, J. Seop Jeong, S. Blackshaw, L.-Z. Jiang, H. Zhu, L. Wu, and Y.-Z. Li, "Identification of New Autoantigens for Primary Biliary Cirrhosis Using Human Proteome Microarrays," *Molecular & Cellular Proteomics*, vol. 11, pp. 669–680, 2012.

[366] S. B. Rothbart, K. Krajewski, B. D. Strahl, and S. M. Fuchs, "Peptide microarrays to interrogate the "histone code"," in *Methods in Enzymology*, vol. 512, pp. 107–135, Academic Press Inc., 2012.

[367] L. C. Szymczak, H.-Y. Kuo, and M. Mrksich, "Peptide Arrays: Development and Application," *Anals of Chemistry*, vol. 90, no. 1, pp. 266–282, 2018.

[368] H. Liu, C. Voss, and S. S. Li, "Using reciprocal protein-peptide array screening to unravel protein interaction networks," in *Methods in Molecular Biology*, vol. 1555, pp. 429–436, Humana

Press Inc., 2017.

[369] J. R. Falsey, M. Renil, S. Park, S. Li, and K. S. Lam, "Peptide and small molecule microarray for high throughput cell adhesion and functional assays," *Bioconjugate Chemistry*, vol. 12, no. 3, pp. 346–353, 2001.

[370] D. Arranz-Solís, S. Pedraza-Díaz, G. Miró, S. Rojo-Montejo, L. Hernández, L. M. Ortega-Mora, and E. Collantes-Fernández, "Tritrichomonas foetus infection in cats with diarrhea from densely housed origins," *Veterinary Parasitology*, vol. 221, pp. 118–122, 2016.

[371] F. F. Loeffler, J. Pfeil, and K. Heiss, "High-density peptide arrays for malaria vaccine development," in *Methods in Molecular Biology*, vol. 1403, pp. 569–582, Humana Press Inc., 2016.

[372] J. Paulo, V. Kadiyala, A. Gaun, and D. Conwell, "Analysis of endoscopic pancreatic function test (ePFT)-collected pancreatic fluid proteins precipitated via ultracentrifugation," *Journal of the Pancreas*, vol. 14, no. 2, 2013.

[373] N. H. Tran, R. Qiao, L. Xin, X. Chen, C. Liu, X. Zhang, B. Shan, A. Ghodsi, and M. Li, "Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry," *Nature Methods*, vol. 16, no. 1, pp. 63–66, 2019.

[374] T. Valikangas, T. Suomi, and L. L. Elo, "A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1344–1355, 2018.

[375] Magellan https://lifesciences.tecan.com/software-magellan (Accessed 06.07.20), "Magellan-Data Analysis Software," 2020.

[376] C. S. Hughes, S. Moggridge, T. Müller, P. H. Sorensen, G. B. Morin, and J. Krijgsveld, "Single-pot, solid-phase-enhanced sample preparation for proteomics experiments," *Nature Protocols*, vol. 14, no. 1, pp. 68–85, 2019.

[377] T. Y. Huang, L. M. Chi, and K. Y. Chien, "Size-exclusion chromatography using reverse-phase columns for protein separation," *Journal of Chromatography A*, vol. 1571, pp. 201–212, 2018.

[378] S. Tyanova, T. Temu, and J. Cox, "The MaxQuant computational platform for mass spectrometry-based shotgun proteomics," *Nature Protocols*, vol. 11, no. 12, pp. 2301–2319, 2016.

[379] S. Tyanova and J. Cox, "Perseus: A bioinformatics platform for integrative analysis of proteomics data in cancer research," *Methods in Molecular Biology*, vol. 1711, pp. 133–148, 2018.

[380] J. D. Bendtsen, H. Nielsen, G. Von Heijne, and S. Brunak, "Improved prediction of signal peptides: SignalP 3.0," *Journal of Molecular Biology*, vol. 340, no. 4, pp. 783–795, 2004.

[381] J. J. Almagro Armenteros, K. D. Tsirigos, C. K. Sonderby, T. N. Petersen, O. Winther, S. Brunak, G. von Heijne, and H. Nielsen, "SignalP 5.0 improves signal peptide predictions using deep neural networks," *Nature Biotechnology*, vol. 37, no. 4, pp. 420–423, 2019.

[382] A. M. Bolger, M. Lohse, and B. Usadel, "Genome analysis Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014.

[383] M.-S. Kim and D. Leahy, "Enzymatic Deglycosylation of Glycoproteins," *Laboratory Methods in Enzymology*, vol. 533, pp. 259–263, 2013.

[384] M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, S. L. Salzberg, and N. Biotechnol, "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads HHS Public Access Author manuscript," *Nature Biotechnol*, vol. 33, no. 3, pp. 290–295, 2015.

[385] PEPerPRINT - https://www.pepperprint.com/ (Accessed 06.07.20), "PEPperPRINT," 2020.

[386] M. E. Ritchie, J. Silver, A. Oshlack, M. Holmes, D. Diyagama, A. Holloway, and G. K. Smyth, "A comparison of background correction methods for two-colour microarrays," *Bioinformatics*, vol. 23, no. 20, pp. 2700–2707, 2007.

[387] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, 2015.

[388] G. K. Smyth, G. K. Smyth, M. Ritchie, N. Thorne, and J. Wettenhall, "LIMMA: linear models for microarray data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor.," *Statistics for Biology and Health*, pp. 397—-420, 2005.

[389] G. K. Smyth and T. Speed, "Normalization of cDNA microarray data," *Methods*, vol. 31, no. 4, pp. 265–273, 2003.

[390] W. Huber, A. Von Heydebreck, H. S. Ultmann, A. Poustka, and M. Vingron, "Variance stabilization applied to microarray data calibration and to the quantification of differential expression," *Bioinformatics*, vol. 18, pp. 96–104, 2002.

[391] G. Casella, "Illustrating empirical Bayes methods," *Chemometrics and Intelligent Laboratory Systems*, vol. 16, no. 2, pp. 107–125, 1992.

[392] T. Liu, H. Zhao, and T. Wang, "An empirical Bayes approach to normalization and differential abundance testing for microbiome data," *BMC Bioinformatics*, vol. 21, no. 1, p. 225, 2020.

[393] T. Vasanthakumar and J. L. Rubinstein, "Structure and Roles of V-type ATPases," *Trends in Biochemical Sciences*, vol. 45, no. 4, pp. 295–307, 2020.

[394] Interpro, "C2 domain (IPR000008) - InterPro entry - InterPro https://www.ebi.ac.uk/interpro/entry/InterPro/IPR000008/ (Accessed 17.08.20)," 2020.

[395] M. K. Dougherty and D. K. Morrison, "Unlocking the code of 14-3-3," *Journal of Cell Science*, vol. 117, no. 10, pp. 1875–1884, 2004.

[396] A. Diaz-Ramos, A. Roig-Borrellas, A. Garcia-Melero, and R. L. Opez-Alemany, "α-Enolase, a Multifunctional Protein: Its Role on Pathophysiological Situations," *Journal of Biomedicine and Biotechnology*, vol. 7, 2012.

[397] K. Krishnan and P. D. Moens, "Structure and functions of profilins," *Biophysical Reviews*, vol. 1, no. 2, pp. 71–81, 2009.

[398] R. Adamo, A. Nilo, B. Castagner, O. Boutureira, F. Berti, and G. J. Bernardes, "Synthetically defined glycoprotein vaccines: Current status and future directions," *Chemical Science*, vol. 4, no. 8, pp. 2995–3008, 2013.

[399] E. Margolin, R. Chapman, A. L. Williamson, E. P. Rybicki, and A. E. Meyers, "Production of complex viral glycoproteins in plants as vaccine immunogens," 2018.

[400] J. Stafkova, P. Rada, D. Meloni, V. Zarsky, T. Smutna, N. Zimmann, K. Harant, P. Pompach, I. Hrdy, and J. Tachezy, "Dynamic secretome of Trichomonas vaginalis: Case study of -amylases," *Molecular and Cellular Proteomics*, vol. 17, no. 2, pp. 304–320, 2018.

[401] A. Rath, S. Choudhury, D. Batra, S. V. Kapre, C. E. Rupprecht, and S. K. Gupta, "DNA vaccine for rabies: Relevance of the trans-membrane domain of the glycoprotein in generating an antibody response," *Virus Research*, vol. 113, no. 2, pp. 143–152, 2005.

[402] R. Kovjazin, I. Volovitz, Y. Daon, T. Vider-Shalit, R. Azran, L. Tsaban, L. Carmon, and Y. Louzoun, "Signal peptides and trans-membrane regions are broadly immunogenic and have high CD8+ T cell epitope densities: Implications for vaccine development," *Molecular Immunology*, vol. 48, no. 8, pp. 1009–1018, 2011.

[403] L. Sun, J. Rush, I. Ghosh, J. R. Maunus, and M. Q. Xu, "Producing peptide arrays for epitope mapping by intein-mediated protein ligation," *BioTechniques*, vol. 37, no. 3, pp. 430–443, 2004.

[404] M. James Francis, "Recent Advances in Vaccine Technologies," *The Veterinary Clinics of North America. Small Animal Practice*, vol. 48, no. 2, 2017.

[405] J. Jeong, C. Park, S. Kim, S. J. Park, I. Kang, K. H. Park, and C. Chae, "Evaluation of the efficacy of a novel porcine circovirus type 2 synthetic peptide vaccine," *Canadian Journal of Veterinary Research*, vol. 82, no. 2, pp. 146–153, 2018.

[406] A. Adams, "Progress, challenges and opportunities in fish vaccine development," *Fish and Shellfish Immunology*, vol. 90, pp. 210–214, 2019.

[407] A. R. Harris, L. Peter, J. Bellis, B. Baum, A. J. Kabla, and G. T. Charras, "Characterizing the mechanics of cultured cell monolayers," *Proceedings of the National Academy of Sciences*

of the United States of America, vol. 109, no. 41, pp. 16449–16454, 2012.

[408] J. S. Dutton, S. S. Hinman, R. Kim, Y. Wang, and N. L. Allbritton, "Primary Cell-Derived Intestinal Models: Recapitulating Physiology," 2019.

[409] B. D. Jett and M. S. Gilmore, "Host-Parasite Interactions in Staphylococcus aureus Keratitis," *DNA and Cell Biology*, vol. 21, no. 5-6, pp. 397–404, 2002.

[410] B. Guggenheim, R. Gmür, J. C. Galicia, P. G. Stathopoulou, M. R. Benakanakere, A. Meier, T. Thurnheer, and D. F. Kinane, "In vitro modeling of host-parasite interactions: the 'subgingival' biofilm challenge of primary human epithelial cells," *BMC Microbiology*, vol. 1, pp. 1–12, 2009.

[411] D. Cardenas, S. Bhalchandra, H. Lamisere, Y. Chen, X.-L. Zeng, S. Ramani, U. C. Karandikar, D. L. Kaplan, M. K. Estes, and H. D. Ward, "Two-and Three-Dimensional Bioengineered Human Intestinal Tissue Models for Cryptosporidium," *Methods in Molecular Biology*, vol. 20, pp. 373–402, 2020.

[412] A. Chatterjee, D. M. Ratner, C. M. Ryan, P. J. Johnson, B. R. O'Keefe, W. E. Secor, D. J. Anderson, P. W. Robbins, and J. Samuelson, "Anti-Retroviral Lectins Have Modest Effects on Adherence of Trichomonas vaginalis to Epithelial Cells In Vitro and on Recovery of Tritrichomonas foetus in a Mouse Vaginal Model," *PLOS ONE*, vol. 10, no. 8, 2015.

[413] M. K. Tolbert, S. H. Stauffer, M. D. Brand, and J. L. Gookin, "Cysteine Protease Activity of Feline Tritrichomonas foetus Promotes Adhesion-Dependent Cytotoxicity to Intestinal Epithelial Cells," *Infection and Immunity*, vol. 7, pp. 2851–9, 2014.

[414] B. N. Singh, J. J. Lucas, D. H. Beach, S. T. Shin, and R. O. Gilbert, "Adhesion of Tritrichomonas foetus to bovine vaginal epithelial cells," *Infection and Immunity*, vol. 67, no. 8, pp. 3847–3854, 1999.

[415] B. N. Singh, "Lipophosphoglycan-like glycoconjugate of Tritrichomonasfoetus and Trichomonas vaginalis," *Molecular and Biochemical Parasitology*, vol. 57, pp. 281–294, 1993.

[416] B. N. Singh, G. R. Hayes, J. J. Lucas, U. Sommer, N. Viseux, E. Mirgorodskaya, R. T. Trifonova, R. R. S. Sassi, C. E. Costello, and R. N. Fichorova, "Structural details and composition of Trichomonas vaginalis lipophosphoglycan in relevance to the epithelial immune function," *Glycoconjugate Journal*, vol. 26, no. 1, pp. 3–17, 2009.

[417] S. Ardalan, B. Craig Lee, and G. E. Garber, "Trichomonas vaginalis: The adhesins AP51 and AP65 bind heme and hemoglobin," *Experimental Parasitology*, vol. 121, no. 4, pp. 300–306, 2009.

[418] M. K. Tolbert, S. H. Stauffer, and J. L. Gookin, "Feline Trichomonas foetus Adgere to Intestinal Epithlium by Receptor-Ligand-Depenent Mechanisms," *Veterinary Parasitology*,

vol. 18, no. 192, pp. 75–82, 2013.

[419] J. F. Alderete and E. Pearlman, "Pathogenic Trichomonas vaginalis cytotoxicity to cell culture monolayers," *The British Journal of Venereal Diseases*, vol. 60, pp. 99–105, 1984.

[420] J. N. Krieger, J. Ravdin, and M. F. Rein, "Contact-Dependent Cytopathogenic Mechanisms of Trichomonas vaginalis," *INFECTION AND IMMUNITY*, pp. 778–786, 1985.

[421] A. Amin, I. Bilic, E. Berger, and M. Hess, "Trichomonas gallinae, in comparison to Tetratrichomonas gallinarum, induces distinctive cytopathogenic effects in tissue cultures," *Veterinary Parasitology*, vol. 186, no. 3-4, pp. 196–206, 2012.

[422] F. F. Pindak, W. A. Gardner, and M. M. De Pindak, "Growth and Cytopathogenicity of Trichomonas vaginalis in Tissue Cultures," Tech. Rep. 4, 1986.

[423] J. F. Alderete, L. Kasmala, E. Metcalfe, and G. E. Garza, "Phenotypic variation and diversity among Trichomonas vaginalis isolates and correlation of phenotype with trichomonal virulence determinants," *Infection and Immunity*, vol. 53, no. 2, pp. 285–293, 1986.

[424] F. C. e.Silva Filho, C. A. Elias, and W. de Souza, "Further studies on the surface charge of various strains of Trichomonas vaginalis and Tritrichomonas foetus," *Cell Biophysics*, vol. 8, no. 3, pp. 161–176, 1986.

[425] B. Da Rocha-Azevedo, M. B. De Melo-Braga, and F. C. E Silva-Filho, "Intra-strain clonal phenotypic variation of Tritrichomonas foetus is related to the cytotoxicity exerted by the parasite to cultured cells," *Parasitology Research*, vol. 95, no. 2, pp. 106–112, 2005.

[426] F. Costa, S. Filho1, and W. De Souza, "The Interaction of Trichomonas vaginalis and Tritrichomonas foetus with Epithelial Cells in Vitro," *Cell Structure and Function*, vol. 13, pp. 301–310, 1988.

[427] K. C. Childers and E. D. Garcin, "Structure/function of the soluble guanylyl cyclase catalytic domain," *Nitric Oxide - Biology and Chemistry*, vol. 77, pp. 53–64, 2018.

[428] J. L. Siqueira-Neto, A. Debnath, L. I. McCall, J. A. Bernatchez, M. Ndao, S. L. Reed, and P. J. Rosenthal, "Cysteine proteases in protozoan parasites," *PLoS Neglected Tropical Diseases*, vol. 12, no. 8, 2018.

[429] M. J. Canova and V. Molle, "Bacterial serine/threonine protein kinases in host-pathogen interactions," *Journal of Biological Chemistry*, vol. 289, no. 14, pp. 9473–9479, 2014.

[430] J. Bugrysheva, B. J. Froehlich, J. A. Freiberg, and J. R. Scott, "Serine/threonine protein kinase Stk is required for virulence, stress response, and penicillin tolerance in Streptococcus pyogenes," *Infection and Immunity*, vol. 79, no. 10, pp. 4201–4209, 2011.

[431] V. Delorme, A. Garcia, X. Cayla, and I. Tardieux, "A role for Toxoplasma gondii type 1 ser/thr protein phosphatase in host cell invasion," *Microbes and Infection*, vol. 4, no. 3,

pp. 271–278, 2002.

[432] J. D. Hemmink, W. Weir, N. D. MacHugh, S. P. Graham, E. Patel, E. Paxton, B. Shiels, P. G. Toye, W. I. Morrison, and R. Pelle, "Limited genetic and antigenic diversity within parasite isolates used in a live vaccine against Theileria parva," *International Journal for Parasitology*, vol. 46, no. 8, pp. 495–506, 2016.

[433] E. Patel, S. Mwaura, H. Kiara, S. Morzaria, A. Peters, and P. Toye, "Production and dose determination of the Infection and Treatment Method (ITM) Muguga cocktail vaccine used to control East Coast fever in cattle," *Ticks and Tick-borne Diseases*, vol. 7, no. 2, pp. 306–314, 2016.

[434] V. Nene, A. Musoke, E. Gobright, and S. Morzaria, "Conservation of the Sporozoite p67 Vaccine Antigen in Cattle-Derived Theileria parva Stocks with Different Cross-Immunity Profiles †," Tech. Rep. 6, 1996.

[435] G. S. Amzati, A. Djikeng, D. O. Odongo, H. Nimpaye, K. P. Sibeko, J.-B. B. Muhigwa, M. Madder, N. Kirschvink, and T. Marcotty, "Genetic and antigenic variation of the bovine tick-borne pathogen Theileria parva in the Great Lakes region of Central Africa," *Parasites Vectors*, vol. 12, no. 588, 2019.

[436] S. A. Gagliano, S. Sengupta, C. Sidore, A. Maschio, F. Cucca, D. Schlessinger, and G. R. Abecasis, "Relative impact of indels versus SNPs on complex disease," *Genetic Epidemiology*, vol. 43, no. 1, pp. 112–117, 2019.

[437] B. S. Shastry, *SNPs: impact on gene function and phenotype.*, vol. 578. Humana Press, Totowa, NJ, 2009.

[438] P. De Wit, M. H. Pespeni, and S. R. Palumbi, "SNP genotyping and population genomics from expressed sequences - current advances and future possibilities," *Molecular Ecology*, vol. 24, no. 10, pp. 2310–2323, 2015.

[439] D. Zhong, C. Koepfli, L. Cui, and G. Yan, "Molecular approaches to determine the multiplicity of Plasmodium infections," *Malaria Journal*, vol. 17, no. 1, pp. 1–9, 2018.

[440] R. Daniels, S. K. Volkman, D. A. Milner, N. Mahesh, D. E. Neafsey, D. J. Park, D. Rosen, E. Angelino, P. C. Sabeti, D. F. Wirth, and R. C. Wiegand, "A general SNP-based molecular barcode for Plasmodium falciparum identification and tracking," *Malaria Journal*, vol. 7, no. 1, p. 223, 2008.

[441] E. Janecek, S. Streichan, and C. Strube, "SNP-based real-time pyrosequencing as a sensitive and specific tool for identification and differentiation of Rickettsia species in Ixodes ricinus ticks," *BMC Infectious Diseases*, vol. 12, p. 1, 2012.

[442] Scitable Nature, "The Information in DNA Determines Cellular Function via

Translation https://www.nature.com/scitable/topicpage/the-information-in-dna-determines-cellular-function-6523228/ (Accessed 01.10.20)," 2020.

[443] S. Kryazhimskiy and J. B. Plotkin, "The Population Genetics of dN/dS," *PLoS Genetics*, vol. 4, no. 12, 2008.

[444] Stanford, "Population Genetics https://plato.stanford.edu/entries/population-genetics/ (Accessed 07.09.20)," 2012.

[445] S. Casillas and A. Barbadilla, "Molecular Population Genetics," *Genetics*, vol. 205, pp. 1003–1035, 2017.

[446] M. Tibayrenc, "Population Genetics of Parasitic Protozoa and other Microorganisms," *Advances in Parasitology*, vol. 36, pp. 47–115, 1995.

[447] A. E. Barry, L. Schultz, C. O. Buckee, and J. C. Reeder, "Contrasting population structures of the genes encoding ten leading vaccine-candidate antigens of the human malaria parasite, Plasmodium falciparum," *PLoS ONE*, vol. 4, no. 12, 2009.

[448] T. Huyse, R. Poulin, and A. Théron, "Speciation in parasites: A population genetics approach," *Trends in Parasitology*, vol. 21, no. 10, pp. 469–475, 2005.

[449] Durvasula, "Interpreting Tajima's D https://arundurvasula.wordpress.com/2015/02/18/interpreting-tajimas-d/ (Accessed 22.10.20)," 2020.

[450] T. S. Korneliussen, I. Moltke, A. Albrechtsen, and R. Nielsen, "Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data," *BMC Bioinformatics*, vol. 14, no. 1, p. 289, 2013.

[451] G. Mcvean, *Natural Selection.* 2002.

[452] T. E. Paulish-Miller, P. Augostini, J. A. Schuyler, W. L. Smith, E. Mordechai, M. E. Adelson, S. E. Gygax, W. E. Secor, and D. W. Hilberta, "Trichomonas vaginalis metronidazole resistance is associated with single nucleotide polymorphisms in the nitroreductase genes ntr4Tv and ntr6Tv," *Antimicrobial Agents and Chemotherapy*, vol. 58, no. 5, pp. 2938–2943, 2014.

[453] A. Mckenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. Depristo, "The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Research*, vol. 20, pp. 1297–1303, 2010.

[454] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo, "From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline," *Current Protocols in Bioinformatics*, vol. 11, no. SUPL.43, p. 11.10.1, 2013.

[455] Simon Andrews, "Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed 07.09.20)," 2010.

[456] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey, "RNA-Seq gene expression estimation with read mapping uncertainty.," *Bioinformatics*, vol. 26, no. 4, pp. 493–500, 2010.

[457] Picard, "Picard Tools - By Broad Institute http://broadinstitute.github.io/picard/ (Accessed 04.09.20)," 2020.

[458] B. Pfeifer, U. Wittelsbürger, S. E. Ramos-Onsins, and M. J. Lercher, "PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R," *Molecular Biology Evolution*, vol. 31, no. 7, pp. 1929–1936, 2014.

[459] B. Pfeifer, "An introduction to the PopGenome package," tech. rep., 2020.

[460] H. G. Le, J. M. Kang, M. Moe, H. Jun, T. L. Thai, J. Lee, M. K. Myint, K. Lin, W. M. Sohn, H. J. Shin, T. S. Kim, and B. K. Na, "Genetic polymorphism and natural selection of circumsporozoite surface protein in Plasmodium falciparum field isolates from Myanmar," *Malaria Journal*, vol. 17, no. 1, p. 361, 2018.

[461] N. Saralamba, M. Mayxay, P. N. Newton, F. Smithuis, F. Nosten, L. Archasuksan, S. Pukrittayakamee, N. J. White, N. P. Day, A. M. Dondorp, and M. Imwong, "Genetic polymorphisms in the circumsporozoite protein of Plasmodium malariae show a geographical bias," *Malaria Journal*, vol. 17, no. 1, p. 269, 2018.

[462] J. Ankarklev, O. Franzén, D. Peirasmaki, J. Jerlström-Hultqvist, M. Lebbad, J. Andersson, B. Andersson, and S. G. Svärd, "Comparative genomic analyses of freshly isolated Giardia intestinalis assemblage A isolates," *BMC Genomics*, vol. 16, no. 1, 2011.

[463] J. Eichler, "Protein glycosylation," *Current Biology*, vol. 29, no. 7, pp. 229–231, 2019.

[464] B. Arey, "The Role of Glycosylation in Receptor Signaling," in *Glycosylation*, InTech, 2012.

[465] L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, and M. J. Sternberg, "The Phyre2 web portal for protein modeling, prediction and analysis," *Nature Protocols*, vol. 10, no. 6, pp. 845–858, 2015.

[466] M. Dalsass, A. Brozzi, D. Medini, and R. Rappuoli, "Comparison of Open-Source Reverse Vaccinology Programs for Bacterial Vaccine Antigen Discovery," *Frontiers in Immunology*, vol. 10, p. 113, 2019.

[467] J. Slapeta, S. Craig, D. Mcdonell, and D. Emery, "Tritrichomonas foetus from domestic cats and cattle are genetically distinct," *Experimental Parasitology*, vol. 126, no. 2, pp. 209–13, 2010.

[468] G-Biosciences, "Recombinant Protein Expression Systems: Pros & Cons

https://info.gbiosciences.com/blog/recombinant-protein-expression-systems-pros-cons (Accessed 09.11.20)," 2020.

[469] Z. Zhang, Y. Li, H. Li, X. Song, Z. Ma, H. Lu, S. Liu, Y. Zhao, M. Tan, S. Wang, and X. Li, "Identification of Toxoplasma Gondii Tyrosine Hydroxylase (TH) Activity and Molecular Immunoprotection against Toxoplasmosis," *Vaccines*, vol. 8, p. 158, 2020.

[470] Z. Zhang, M. Elizbeth Alvarez Sanchez, X. Li, Y. Li, S. Wang, L. Hao, Y. Zhu, H. Li, X. Song, Y. Duan, Y. Sang, and P. Wu, "The Molecular Characterization and Immunity Identification of Trichomonas vaginalis Adhesion Protein 33 (AP33)," *Frontiers in Microbiology*, vol. 11, p. 1433, 2020.

[471] R. M. Bill, "Playing catch-up with Escherichia coli: using yeast to increase success rates in recombinant protein production experiments," *Frontiers in Microbiology*, vol. 5, no. MAR, p. 85, 2014.

[472] A. Milton Vieira Gomes, T. Souza Carmo, L. Silva Carvalho, F. I. Mendonça Bahia, and N. Skorupa Parachin, "microorganisms Comparison of Yeasts as Hosts for Recombinant Protein Production," *Microorganisms*, vol. 6, no. 38, 2018.

[473] E. Lisowska, "The role of glycosylation in protein antigenic properties," *Cellular and Molecular Life Sciences*, vol. 59, no. 3, pp. 445–455, 2002.

[474] M. A. Wolfert, G.-J. Boons, N. Chem, and B. Author, "Adaptive immune activation: glycosylation does matter," *Nature Chemical Biology*, vol. 9, no. 12, pp. 776–784, 2013.

[475] J. B. Goh and S. K. Ng, "Impact of host cell line choice on glycan profile," *Critical Reviews in Biotechnology*, vol. 38, no. 6, pp. 851–867, 2018.

[476] C. L. Morton and P. M. Potter, "Comparison of Escherichia coli, Saccharomyces cerevisiae, Pichia pastoris, Spodoptera frugiperda, and COS7 cells for recombinant gene expression: Application to a rabbit liver carboxylesterase," *Applied Biochemistry and Biotechnology*, vol. 16, no. 3, pp. 193–202, 2000.

[477] E. M. Meulenbroek, D. Wouters, and S. Zeerleder, "Methods for Quantitative Detection of Antibody-induced Complement Activation on Red Blood Cells," *Journal of Visualised Experiments*, no. 83, p. 51161, 2014.

[478] G. Stasilojc, A. Felberg, A. Urban, D. Kowalska, S. Ma, A. M. Blom, J. Lundin, A. Österborg, and M. Okrój, "Calcein release assay as a method for monitoring serum complement activity during monoclonal antibody therapy in patients with B-cell malignancies," *Journal of Immunological Methods*, vol. 476, p. 112675, 2020.

[479] M. Costabile, "Measuring the 50% Haemolytic Complement (CH50) Activity of Serum," *Journal of Visualised Experiments*, vol. 37, 2010.

[480] H. P. Weerts, J. Yu, R. W. Kaminski, and M. H. Nahm, "A High-throughput Shigella-specific Bactericidal Assay," *Journal of Visualised Experiments*, no. 144, p. 59164, 2019.

[481] C. Dando, K. E. Gabriel, J. S. Remington, and S. F. Parmley, "Simple and efficient method for measuring anti-toxoplasma immunoglobulin antibodies in human sera using complement-mediated lysis of transgenic tachyzoites expressing β-galactosidase," *Journal of Clinical Microbiology*, vol. 39, no. 6, pp. 2122–2125, 2001.

[482] B. Meissner, T. Rogalski, R. Viveiros, A. Warner, and L. Plastino, "Determining the Sub-Cellular Localization of Proteins within Caenorhabditis elegans Body Wall Muscle," *PLoS ONE*, vol. 6, no. 5, 2011.

[483] L. I. FitzGerald, L. Aurelio, M. Chen, D. Yuen, J. J. Rennick, B. Graham, and A. P. Johnston, "A molecular sensor to quantify the localization of proteins, DNA and nanoparticles in cells," *Nature Communications*, vol. 11, no. 1, 2020.

[484] R. N. Besingi and P. L. Clark, "Extracellular Protease Digestion to Evaluate Membrane Protein Cell Surface Localization," *Nature Protocols*, vol. 10, no. 12, pp. 2074–2080, 2015.

[485] T. S. Edrington, G. H. Loneragan, J. Hill, K. J. Genovese, D. M. Brichta-Harhay, R. L. Farrow, N. A. Krueger, T. R. Callaway, R. C. Anderson, and D. J. Nisbet, "Development of challenge models to evaluate the efficacy of a vaccine to reduce carriage of salmonella in peripheral lymph nodes of cattle3," *Journal of Food Protection*, vol. 76, no. 7, pp. 1259–1263, 2013.

[486] G. A. Albanese, L. R. Tensa, E. J. Aston, D. A. Hilt, and B. J. Jordan, "Evaluation of a coccidia vaccine using spray and gel applications," *Poultry Science*, vol. 97, pp. 1544–1553, 2018.

[487] M. M. Cooper, C. Loiseau, J. S. McCarthy, and D. L. Doolan, "Human challenge models: tools to accelerate the development of malaria vaccines," 2019.

[488] K. Matuschewski, "Murine infection models forvaccine development: The malaria example," *Human Vaccines and Immunotherapeutics*, vol. 9, no. 3, pp. 450–456, 2013.

[489] B. M. Buddle, N. A. Parlane, D. N. Wedlock, and A. Heiser, "Overview of Vaccination Trials for Control of Tuberculosis in Cattle, Wildlife and Humans," *Transboundary and Emerging Diseases*, vol. 60, pp. 136–146, 2013.

[490] M. Z. Farooq, S. Sattar, H. Afroz, and A. Rasool, "A novel approach for development, standardization, and safety testing of enriched alum-precipitated vaccine against hemorrhagic septicemia in different breeds of cattle," *Tropical Animal Health and Production*, vol. 51, no. 4, pp. 839–845, 2019.

Appendices

.1 Chapter 2 Appendix

Transcriptome Condition	No. of Reads	Overall Alignment Rate (%)
Control 1	67634690	96.04
Control 2	15521536	97.22
Control 3	14386258	97.12
Control 4	14490527	97.39
Control 5	18342427	97.38
Colchicine 1	53132714	93.16
Colchicine 2	39965884	5.29
Colchicine 3	23430572	61.03
Low Temperature 1	67107494	91.86
Low Temperature 2	54584264	91.94
Low Temperature 3	58026576	90.35
High Temperature (42°C) 1	17228591	97.36
High Temperature (42°C) 2	15035565	97.45
High Temperature (42°C) 3	16267485	96.52
High Temperature (46°C) 1	14733629	96.45
High Temperature (46°C) 2	17161827	96.89
High Temperature (46°C) 3	13449004	96.60
High pH 1	17422478	96.29
High pH 2	16157854	93.04
Low pH 1	19597891	97.46
Low pH 2	15681152	92.47
Low pH 3	13862147	95.64
Oxidative Stress (500) 1	23622541	96.76
Oxidative Stress (500) 2	13969196	70.64
Oxidative Stress (300) 1	21351760	96.97
Oxidative Stress (300) 2	20516795	92.32
Glucose-free media 1	36099598	94.94
Glucose-free media 2	32593301	95.72
Glucose-free media 3	30682198	94.84
Tryptone-free media 1	16999365	90.33
Tryptone-free media 2	37760351	95.49
Serum-free media 1	31774637	95.13
Serum-free media 2	36355202	95.66
Serum-free media 3	36948206	96.03

Table A1: Mapping statistics for *T. foetus* transcriptomes.The initial sequences were sequenced using Hiseq at the Liverpool CGR and statistics calculated using the HiSat2 package on the Galaxy Server. Where numbers '1, 2 and 3' refer to replicates.

Low Temperature Upregulated		
Gene ID	Fold change	Gene Product
TTF74316	14.34	Hypothetical
TTF74318	12.56	Hypothetical
TTF74053	12.45	Hypothetical
TTF74317	11.88	Hypothetical
TTF74052	11.83	Hypothetical
TTF74054	11.65	Hypothetical
TTF40717	10.93	N-acetyltransferase
TTF42099	10.63	N-acetyltransferase
TTF44477	10.55	Hypothetical
TTF09898	10.16	Hypothetical
Low Temperature Downregulated		
Gene ID	Fold change	Gene Product
TTF81834	-9.10	Transposable element t3 transposase
TTF24549	-7.79	Myb-like DNA-binding domain containing protein
TTF38890	-7.50	Hypothetical
TTF33346	-7.36	Flavodoxin-like fold family protein
TTF33132	-7.35	Transposable element t3 transposase
TTF06492	-7.15	Hypothetical
TTF55653	-7.05	Hypothetical
TTF12964	-6.90	NA
TTF33007	-6.87	Hypothetical
TTF59816	-6.84	NA

Table A2: Differential expression of low temperature induced pseudocysts relative to the control Trophozoites, for both up and downregulated genes the top 10 with the highest fold change are shown.

High Temperature 42°C Upregulated		
Gene ID	Fold change	Gene Product
TTF08541	5.43	Calcium Translocating P-type ATPase
TTF23546	5.25	Hypothetical
TTF72038	5.23	Hypothetical
TTF81061	5.07	NA
TTF20278	4.99	NA
TTF23324	4.95	GNAT family N-acetyltransferase
TTF36525	4.83	ABC transporter ATP-binding domain
TTF01010	4.80	Helicase carboxy terminal domain
TTF70954	4.63	Hypothetical
TTF39485	4.53	Hypothetical
High Temperature 42°C Downregulated		
Gene ID	Fold change	Gene Product
TTF54491	-29.65	Glycosyl hydrolase
TTF33467	-28.33	CAMK family protein
TTF35720	-22.35	Hypothetical
TTF54196	-25.23	Skin secretory protein xp2-like
TTF81834	-24.70	Transposable element
TTF00196	-24.55	NA
TTF67826	-24.29	NA
TTF23815	-13.46	Cysteine protease
TTF23813	-14.43	Cysteine protease
TTF54492	-12.52	Glycosyl hydrolase

Table A3: Differential expression of High Temperature (42°C) exposed *T. foetus* cells compared to control trophozoites, for both up and downregulated genes the top 10 with the highest fold change are shown.

High Temperature 46°C Upregulated		
Gene ID	Fold change	Gene Product
TTF74570	6.02	Hypothetical
TTF57624	5.92	ABC Transporter
TTF08569	5.90	Phosphoglycerol transferase
TTF06433	5.81	Hypothetical
TTF46339	5.56	Hypothetical
TTF22177	5.63	Hypothetical
TTF23557	5.54	Myb-like DNA-binding domain containing protein
TTF11466	5.54	Major Facilitator superfamily transporter
TTF44911	5.41	Histone-lycine N-methyl transferase
TTF07238	4.04	Hypothetical
High Temperature 46°C Downregulated		
Gene ID	Fold change	Gene Product
TTF23815	-14.91	Cysteine protease
TTF54491	-14.05	Glycosyl hydrolase
TTF33467	-13.10	CAMK family protein
TTF32002	-12.74	Myb-like DNA-binding domain containing protein
TTF23818	-12.12	Cysteine protease
TTF54492	-11.73	Glycosyl hydrolase
TTF54493	-11.73	Hypothetical
TTF23815	-11.43	Cysteine protease
TTF81875	-11.01	Cytoplasmic heat shock protein
TTF15376	-10.94	SH3 domain containing protein

Table A4: Differential expression of High Temperature (46°C) exposed *T. foetus* cells compared to control trophozoites, for both up and downregulated genes the top 10 with the highest fold change are shown.

Oxidative stress 300mM Upregulated		
Gene ID	Fold change	Gene Product
TTF67650	7.01	HECT E5 ubiquitin ligase
TTF26415	6.75	Hypothetical
TTF82770	6.63	CAMK family protein
TTF30026	6.55	Protein phosphotase
TTF10206	6.52	Keratinocyte proline-rich protein like
TTF58873	6.46	Myb-like DNA-binding domain containing protein
TTF37785	6.43	Hypothetical
TTF51385	6.41	Hypothetical
TTF55240	6.41	Major facilitatory superfamily
TTF82201	6.38	Hypothetical
Oxidative Stress 300mM Downregulated		
Gene ID	Fold change	Gene Product
TTF54196	-7.18	Skin secretory protein xp2-like
TTF54204	-6.92	NA
TTF17855	-6.67	NA
TTF42969	-5.54	Acetylhydrolase
TTF36815	-5.51	NA
TTF51460	-5.17	NA
TTF16293	-5.06	Hypothetical
TTF32590	-4.88	Hypothetical
TTF54061	-4.77	TCCD-inducible-PARP-like domain
TTF71281	-4.65	Hypothetical

Table A5: Differential expression of *T foetus* cells exposed to oxidative stress of 300mM hydrogen peroxide compared to control trophozoites, for both up and downregulated genes the top 10 with the highest fold change are shown.

Oxidative stress 500mM Upregulated		
Gene ID	Fold change	Gene Product
TTF32010	8.59	Hypothetical
TTF09943	7.92	Dyenin heavy chai family protein
TTF43762	7.33	Hypothetical
TTF14719	7.18	Hypothetical
TTF39941	7.18	Hypothetical
TTF62578	7.08	Hypothetical
TTF32900	7.06	Hypothetical
TTF10038	7.03	E3 ubiquitin protein
TTF38527	7.01	Hypothetical
TTF46338	6.93	Hypothetical
Oxidative Stress 500mM Downregulated		
Gene ID	Fold change	Gene Product
TTF54196	-8.36	Skin secretory protein xp2-like
TTF43713	-7.36	Adenine-specific methyltransferase
TTF51074	-7.35	Hypothetical
TTF67629	-7.35	Hypothetical
TTF60423	-7.27	Adenine-specific methyltransferase
TTF02258	-7.24	Hypothetical
TTF49034	-7.17	Integrase core domain containing protein
TTF09544	-7.08	Hypothetical
TTF75062	-6.96	DNA polmerase type B
TTF26956	-6.72	Histone lysine N-methyltransferase

Table A6: Differential expression of *T. foetus* cells exposed to oxidative stress of 500 mM hydrogen peroxide compared to control trophozoites, for both up and downregulated genes the top 10 with the highest fold change are shown.

High pH Upregulated		
Gene ID	Fold change	Gene Product
TTF37964	7.99	NA
TTF53975	7.69	Hypothetical
TTF62758	7.55	Hypothetical
TTF31470	7.54	Histidine acid phosphatase
TTF44346	7.44	GNAt family N-acetyltransferase
TTF22638	7.25	Hypothetical
TTF84400	7.20	Hypothetical
TTF37965	7.18	Exportin 1
TTF25039	7.18	Major facilitator superfamily transporter
TTF37669	7.08	Hypothetical
High pH Downregulated		
Gene ID	Fold change	Gene Product
TTF39285	-12.56	Hypothetical
TTF54196	-12.19	Skin secretory protein xp2-like
TTF38866	-11.60	N-acetyltransferase
TTF18006	-11.50	Myb-like DNA-binding domain containing protein
TTF74883	-11.45	Myb-like DNA-binding domain containing protein
TTF79591	-11.17	CAMK family protein
TTF79485	-10.44	Myb-like DNA-binding domain containing protein
TTF37949	-10.21	NA
TTF65543	-10.17	Hypothetical
TTF29000	-10.14	Myb-like DNA-binding domain containing protein

Table A7: Differential expression of *T. foetus* cells exposed to high pH stress compared to control trophozoites, for both up and downregulated genes the top 10 with the highest fold change are shown.

Low pH Upregulated		
Gene ID	Fold change	Gene Product
TTF48041	5.95	Kerotinyte proline-rich repeat
TTF45031	5.45	NA
TTF63369	5.15	NA
TTF11808	4.97	Hypothetical
TTF25260	4.95	NA
TTF24809	4.89	Hypothetical
TTF13772	4.81	NA
TTF05255	4.80	Leucine-rich-repeat
TTF09507	4.76	Hypothetical
TTF19547	4.68	Hypothetical
Low pH Downregulated		
Gene ID	Fold change	Gene Product
TTF54198	-9.92	Skin secretory protein xp2-like
TTF26270	-6.62	NLI-Interaction factor-like phosphatase
TTF62269	-6.15	NA
TTF59396	-6.00	NA
TTF54198	-5.88	Skin secretory protein xp2-like
TTF65164	-5.84	Hypothetical
TTF76380	-5.76	Hypothetical
TTF73685	-5.55	Ser/Thr Protein phosphatase
TTF54200	-5.32	Skin secretory protein
TTF14152	-5.13	Protein phophatase

Table A8: Differential expression of *T. foetus* cells exposed to low pH stress, compared to control trophozoites, for both up and downregulated genes the top 10 with the highest fold change are shown.

Tryptone free Upregulated		
Gene ID	Fold change	Gene Product
TTF54196	14.87	Skin secretory protein xp2-like
TTF54200	11.04	Skin secretory protein xp2-like
TTF82335	10.39	Hypothetical
TTF54198	9.18	Skin secretory protein xp2-like
TTF29736	9.04	NA
TTF54204	8.33	NA
TTF54203	7.97	Skin secretory protein xp2-like
TTF10811	7.58	NA
TTF33899	7.38	Cation efflux protein
TTF67560	7.10	NA
Tryptone free Downregulated		
Gene ID	Fold change	Gene Product
TTF03367	-10.52	RND transporter protein
TTF17133	-10.44	Ras-family protein
TTF65665	-10.23	NA
TTF03369	-10.18	Hypothetical
TTF31558	-10.09	Neimann-pick C1-like domain containing protein
TTF65674	-10.00	Hypothetical
TTF03368	-9.59	Hypothetical
TTF81870	-9.50	Hypothetical
TTF41168	-9.45	Rhodanese-like domain
TTF80844	-9.14	Hypothetical

Table A9: Differential expression of *T. foetus* cells grown in media with no tryptone added compared to control trophozoites grown in standard Diamond media, for both up and downregulated genes the top 10 with the highest fold change are shown.

Serum free Upregulated		
Gene ID	Fold change	Gene Product
TTF54196	13.06	Skin secretory protein xp2-like
TTF10811	10.84	NA
TTF54200	9.76	Skin secretory protein xp2-like
TTF82335	9.10	Hypothetical
TTF61610	8.57	NA
TTF27394	8.46	ABC transporter
TTF14418	8.35	Hypothetical
TTF58441	8.13	Myb-like DNA-binding domain containing protein
TTF49081	8.05	Hypothetical
TTF79613	7.98	Myb-like DNA-binding domain containing protein
Serum free Downregulated		
Gene ID	Fold change	Gene Product
TTF48543	-7.12	Ras related protein
TTF52850	-6.15	DNA polymerase
TTF34502	-5.43	Comm domain containing protein
TTF62451	-5.43	Hypothetical
TTF78238	-5.36	uDENN domain containing protein
TTF03377	-5.33	Phosphotidylinositol phosphodiesterase
TTF28511	-5.27	NA
TTF66345	-5.18	NA
TTF69058	-5.13	Major facillitator superfamily transporter
TTF28773	-5.11	Hypothetical

Table A10: Differential expression of *T. foetus* cells in media with no serum added compared to control trophozoites grown in standard Diamond media, for both up and downregulated genes the top 10 with the highest fold change are shown.

Glucose free Upregulated		
Gene ID	Fold change	Gene Product
TTF54196	15.53	Skin secretory protein xp2-like
TTF10811	14.61	Sperm motility kinase
TTF54200	14.42	Aminotransferase
TTF82335	14.04	Glycine dehydrogenase
TTF61610	12.87	Aminomethyl transferring glycine dehydrogenase
TTF27394	12.81	Hypothetical
TTF14418	12.75	Aminomethyl transferring glycine dehydrogenase
TTF58441	12.29	Hypothetical
TTF49081	12.24	Myb-like DNA-binding domain containing protein
TTF79613	12.10	Glycine clevage system t protein
Glucose free Downregulated		
Gene ID	Fold change	Gene Product
TTF31558	-10.56	Niemann pick c-like protein 1
TTF03367	-9.37	RND family protein
TTF17133	-9.00	Ras family protein
TTF03368	-8.99	Hypothetical
TTF19166	-8.58	Serine-protease inhibitor family protein
TTF53714	-8.43	Acetyltransferase family protein
TTF62567	-8.19	PQ loop repeat family
TTF62049	-8.17	PQ loop repeat family
TTF80844	-7.60	Hypothetical
TTF62568	-7.56	Hypothetical

Table A11: Differential expression of *T. foetus* cells in media with no serum added compared to control trophozoites grown in standard Diamond media, for both up and downregulated genes the top 10 with the highest fold change are shown.

.2 Chapter 3 Appendix

Accession	Trophozoite	Pseudocyst	Description
TTF30249	1	2.54	glycosyl hydrolase
TTF48473	1	2.22	60S ribosomal protein L18a
TTF37123	1	2.17	40S ribosomal protein S3a
TTF00247	1	2.08	40S ribosomal protein S15a
TTF21423	1	2.08	40S ribosomal protein S15a
TTF00248	1	2.08	40S ribosomal protein S15a
TTF10548	1	2.08	40S ribosomal protein S15a
TTF00249	1	2.08	40S ribosomal protein S15a
TTF00246	1	2.08	40S ribosomal protein S15a
TTF05687	1	2.06	Coatomer subunit beta
TTF37642	1	1.99	hypothetical protein TRFO_20197
TTF05663	1	1.97	archaeal ribosomal protein S17P
TTF43167	1	1.95	hypothetical protein TRFO_16532
TTF00060	1	1.94	ribosomal protein L27e
TTF51775	1	1.94	ribosomal protein L27e
TTF01798	1	1.91	cAMP-dependent protein kinase type I-alpha
TTF32301	1	1.9	hypothetical protein TRFO_23773
TTF23412	1	1.9	putative C2 domain containing protein
TTF47374	1	1.88	ribosomal protein S3
TTF79711	1	1.88	ribosomal protein S3
TTF79900	1	1.88	ribosomal protein S3
TTF79814	1	1.88	ribosomal protein S3
TTF45174	1	1.88	ribosomal protein S3
TTF20455	1	1.88	kelch-like protein 24
TTF28444	1	1.86	40S ribosomal protein S3a
TTF05650	1	1.86	40S ribosomal protein S3a
TTF37554	1	1.79	ABC transporter E family member 2
TTF23491	1	1.79	ABC transporter E family member 2
TTF01496	1	1.76	clathrin heavy chain
TTF32482	1	1.72	signal recognition particle subunit SRP72-like
TTF00547	1	1.71	hypothetical protein TRFO_18052

TTF78075	1	1.71	hypothetical protein TRFO_37894
TTF74267	1	1.69	AMP-binding enzyme family protein
TTF62595	1	1.69	hypothetical protein TRFO_05477
TTF68935	1	1.64	oxidoreductase
TTF73129	1	1.63	GTP-binding protein SAR1A
TTF01232	1	1.59	DUF1846 domain-containing protein
TTF00552	1	1.59	ankyrin repeat protein
TTF70519	1	1.58	Rho family GTpase
TTF48213	1	1.58	ribosomal protein L21e
TTF16749	1	1.58	ribosomal protein L21e
TTF07112	1	1.58	60S ribosomal protein L21
TTF04577	1	1.58	60S ribosomal protein L21
TTF66543	1	1.58	cysteine desulfurase
TTF21484	1	1.57	coatomer subunit gamma
TTF14082	1	1.56	Elongation factor 2
TTF74567	1	1.56	Phosphoglucomutase phosphomannomutase
TTF30640	1	1.55	Glycogen debranching enzyme
TTF45115	1	1.55	elongator complex protein 1
TTF26436	1	1.54	sugar nucleotide epimerase
TTF35736	1	1.51	60S ribosomal protein L9
TTF01791	1	1.5	Copine family protein
TTF28341	1	1.5	hypothetical protein
TTF41143	1	0.67	hypothetical protein TRFO_12516
TTF52086	1	0.65	kinetoplast-associated protein
TTF37126	1	0.65	charged multivesicular body protein 3
TTF37125	1	0.65	predicted protein
TTF39902	1	0.65	hypothetical protein TRFO_10466
TTF42678	1	0.65	endoplasmic reticulum-Golgi intermediate compartment
TTF65056	1	0.64	hypothetical protein TRFO_35633
TTF31986	1	0.64	hypothetical protein TRFO_26034
TTF73657	1	0.64	Centrin-3
TTF44549	1	0.63	MBL fold metallo-hydrolase
TTF24016	1	0.63	histidine acid phosphatase
TTF07839	1	0.63	hypothetical protein TRFO_23646

TTF20851	1	0.62	Clan SC
TTF63668	1	0.62	hypothetical protein TRFO_03267
TTF23702	1	0.62	14-3-3 protein
TTF04579	1	0.62	Transmembrane amino acid transporter protein
TTF45738	1	0.62	hypothetical protein TRFO_42261
TTF54599	1	0.61	hypothetical protein TRFO_13511
TTF43209	1	0.61	hypothetical protein TRFO_39443
TTF22383	1	0.61	hypothetical protein TRFO_30043
TTF35926	1	0.59	histidine acid phosphatase
TTF58541	1	0.59	hypothetical protein TRFO_00866
TTF08565	1	0.59	hypothetical protein TRFO_20073
TTF58329	1	0.59	Prefoldin subunit 1
TTF50578	1	0.59	vacuolar proton translocating ATPase
TTF74681	1	0.59	inner centromere protein
TTF45739	1	0.59	hypothetical protein
TTF60517	1	0.58	hypothetical protein TRFO_29767
TTF68546	1	0.58	C2 domain containing protein
TTF39353	1	0.58	NAD synthetase
TTF83994	1	0.58	hypothetical protein TRFO_08132
TTF22382	1	0.58	histidine acid phosphatase
TTF40175	1	0.57	Laminin A family protein
TTF58150	1	0.56	Clan CA
TTF58433	1	0.56	Clan CA
TTF15102	1	0.56	hypothetical protein TRFO_34808
TTF54148	1	0.56	hypothetical protein TRFO_38207
TTF00659	1	0.56	Inositol polyphosphate kinase family protein
TTF42565	1	0.55	Immuno-dominant variable surface antigen-like
TTF70906	1	0.55	hypothetical protein TRFO_12915
TTF67248	1	0.55	coiled-coil domain-containing protein
TTF17368	1	0.55	hypothetical protein TRFO_07600
TTF21472	1	0.55	hypothetical protein TRFO_31205
TTF81703	1	0.54	V-type ATPase 116kDa subunit family protein
TTF53390	1	0.54	predicted protein
TTF62573	1	0.54	hypothetical protein TRFO_05497

TTF36533	1	0.53	Uncharacterised protein
TTF58505	1	0.53	hypothetical protein TRFO_00819
TTF10766	1	0.52	C2 domain containing protein
TTF42563	1	0.52	Immuno-dominant variable surface antigen-like
TTF43593	1	0.51	Clan MH
TTF41233	1	0.51	cathepsin O-like
TTF39115	1	0.51	cell division control protein 31
TTF11319	1	0.51	hypothetical protein TRFO_23477
TTF61143	1	0.5	hypothetical protein TRFO_19729
TTF44089	1	0.5	hypothetical protein TRFO_01273
TTF40594	1	0.48	hydrogenosomal membrane protein 31 precursor
TTF25098	1	0.48	papain family cysteine protease domain containing protein
TTF54808	1	0.48	CAMK family protein kinase
TTF25320	1	0.47	hypothetical protein TRFO_20099
TTF15373	1	0.47	hypothetical protein
TTF59087	1	0.47	hypothetical protein TRFO_05193
TTF54959	1	0.47	hypothetical protein TRFO_14709
TTF43481	1	0.47	hypothetical protein TRFO_10329
TTF01106	1	0.46	hypothetical protein TRFO_08327
TTF55788	1	0.46	hypothetical protein TRFO_17592
TTF81917	1	0.46	ATPase
TTF63245	1	0.46	Clan SB
TTF21701	1	0.45	hypothetical protein TRFO_40634
TTF62655	1	0.45	putative centrin
TTF68155	1	0.45	hypothetical protein TRFO_40979
TTF56506	1	0.44	major antigen-like isoform X1
TTF11427	1	0.43	hypothetical protein TRFO_06752
TTF12145	1	0.43	proliferating cell nuclear antigen
TTF08567	1	0.42	hypothetical protein TRFO_20073
TTF00657	1	0.42	hypothetical protein TRFO_01490
TTF15093	1	0.42	—NA—
TTF12585	1	0.41	hypothetical protein TRFO_38671
TTF13971	1	0.4	kinetoplast-associated protein
TTF67500	1	0.39	Clan SC

TTF36554	1	0.36	major facilitator superfamily transporter
TTF80961	1	0.36	syntaxin-related protein
TTF81042	1	0.31	putative folate-biopterin transporter 2
TTF81043	1	0.31	leishmanolysin family protein

Table B1: Ratio of intensities of TMT labelled pseudocysts relative to trophozoites. When the ratio is higher than 1 in the pseudocyst sample, there is significantly higher intensity of the protein. Where the ratio is less than 1, there is significantly higher intensity of the protein in the trophozoite sample.

Protein ID	Product	Fold Change
TTF01106	hypothetical protein TRFO_08327	36.43962745
TTF44954	60s ribosomal protein l23a	24.64628994
TTF56658	hypothetical protein TRFO_31205	21.66641155
TTF32797	ATP-dependent RNA helicase eIF4A	16.29165891
TTF08567	hypothetical protein TRFO_20073	11.66314474
TTF01817	40S ribosomal protein S2	11.55148153
TTF23865	geranylgeranyl transferase type-2 subunit beta-like	10.90461661
TTF21701	hypothetical protein TRFO_40634	9.520755849
TTF35844	ribosomal protein L32 putative	7.915135316
TTF09389	60S ribosomal protein L37a	7.617905076
TTF13481	Histone H4	6.471337116
TTF20135	ribosomal protein L27e	6.119031455
TTF49479	ribosomal protein S13p S18e	5.348171121
TTF41309	C2 domain containing protein	5.236961108
TTF32892	ribosomal protein L24e	5.103441195
TTF53391	40S ribosomal protein S13	4.650094756
TTF78555	ribosomal protein L35Ae	4.567931793
TTF04577	60S ribosomal protein L21	4.393391297
TTF83969	60S ribosomal protein L23	3.930978844
TTF55912	ribosomal protein L13e	3.881543321
TTF00609	60S ribosomal protein L37	3.801548523
TTF80890	alpha tubulin	3.621496719
TTF49944	40S ribosomal protein S16	3.600275778
TTF59123	Hydrogenase Fe-only	3.365426696
TTF41687	RNA-binding protein	3.356894016
TTF26810	C2 domain containing protein	3.296468656
TTF44008	AMP-binding enzyme family protein	3.018174385
TTF15811	60S ribosomal protein L14	2.928534836
TTF20747	ribosomal protein S14	2.869292992
TTF81253	ribosomal protein L7Ae	2.777875702
TTF20525	ribosomal protein L14	2.591067847
TTF03379	RAC GTPase putative	2.546520935

TTF79182	Tudor domain containing protein	2.53090186
TTF30587	F420-0–gamma-glutamyl ligase	2.492632666
TTF01496	clathrin heavy chain	2.473896323
TTF81691	78 kDa glucose-regulated protein	2.40220795
TTF39123	40S ribosomal protein S7	2.397611797
TTF37123	40S ribosomal protein S3a	2.361080861
TTF07843	beta tubulin	2.231087242
TTF50642	40S ribosomal protein S4 X isoform	2.230788366
TTF02180	40S ribosomal protein S8	2.224038174
TTF41702	Immuno-dominant variable surface antigen-like	2.212639449
TTF40080	40S ribosomal protein S9-1	2.060041442

Table B2: Fold change of biotinylated *T. foetus* biotinylated trophozoites relative to low temperature induced pseudocysts . All elutions and all experimental trials for each life stage were combined then compared. The samples had all been run on the Orbitrap mass spectrometer using label free m/s and the results were analysed in Peaks [373]. All differentially abundant proteins had a fold change >2 a $P< 0.05$.

Protein ID	Product	Fold Change
TTF21425	alanyl-tRNA synthetase	74.44341287
TTF21911	6-phosphogluconate dehydrogenase	43.52033493
TTF02041	Glucokinase 1	29.43518205
TTF24249	OsmC family peroxiredoxin	25.74577122
TTF64999	NADP-specific glutamate dehydrogenase	23.22239141
TTF10541	Rubrerythrin family protein	16.6191032
TTF40382	WD40 domain containing protein	16.40132548
TTF11427	hypothetical protein TRFO_06752	14.9687056
TTF22398	hydrogenase assembly factor	14.22965026
TTF06177	glutamate decarboxylase	13.25025351
TTF44855	hypothetical protein TRFO_30443	9.578736602
TTF33068	protein disulfide-isomerase domain	8.988597843
TTF15804	T-complex protein 1 subunit eta	8.062109075
TTF10766	C2 domain containing protein	7.452111345
TTF28277	V-type proton ATPase subunit H	7.279193768
TTF81588	thioredoxin peroxidase	6.233543893
TTF20914	hypothetical protein TRFO_06622	6.005153428
TTF26427	pyruvate ferredoxin flavodoxin oxidoreductase	5.801389649
TTF36147	sugar O-acetyltransferase	5.658459359
TTF01062	hydrogenosomal membrane protein 31 precursor	5.441286178
TTF39122	Glucokinase 1	5.073289902
TTF23521	Catalase	5.061837325
TTF30749	L-fucose isomerase	4.944426989
TTF23491	ABC transporter E family member 2	4.646384663
TTF47666	myo-inositol-1-phosphate synthase-like protein	4.635849802
TTF66999	elongation factor 1-beta	4.468929758
TTF26391	Transaldolase	4.439203402
TTF29022	Rab8 family GTPase	4.364227719
TTF82156	30S ribosomal protein S19e	4.334407888
TTF22695	Thioredoxin family protein	4.167570561
TTF81884	vacuolar protein sorting-associated protein 35	4.056338948
TTF37447	actin putative	3.916002153

TTF35950	MBL fold metallo-hydrolase	3.910664481
TTF55099	glycylpeptide N-tetradecanoyltransferase 1	3.828385724
TTF79971	BTB POZ domain containing protein	3.756098544
TTF65227	adhesin AP65-1 precursor	3.687693138
TTF00205	T-complex protein 1 subunit alpha	3.60655547
TTF61619	Glucosamine-6-phosphate deaminase	3.547086707
TTF55377	Aldose 1-epimerase	3.502384256
TTF82220	galactose mutarotase	3.502384256
TTF40594	hydrogenosomal membrane protein 31 precursor	3.439136548
TTF07837	putative Arp2 3	3.416055378
TTF81436	thioredoxin peroxidase	3.366773783
TTF51797	cytosolic malate dehydrogenase 1	3.316542701
TTF30959	V-type proton ATPase subunit E	3.265933769
TTF24218	Dynamin central region family protein	3.261656228
TTF36731	hypothetical protein TRFO_08672	3.234768945
TTF62416	ribose-phosphate pyrophosphokinase putative	3.186811392
TTF08207	TPR Domain containing protein	3.092873245
TTF73388	Actin-related protein 2 3 complex subunit 3	3.020048743
TTF77415	AMP-binding enzyme family protein	2.959804154
TTF08733	Glucosamine-6-phosphate deaminase	2.734008587
TTF22965	Malate dehydrogenase	2.723376697
TTF44549	MBL fold metallo-hydrolase	2.722754715
TTF31424	chaperonin GroEL	2.717108135
TTF64548	sarcoplasmic-endoplasmic reticulum calcium ATPase	2.696320314
TTF32664	F-actin capping protein alpha subunit	2.581813293
TTF20133	isoleucine–tRNA ligase cytoplasmic	2.559204573
TTF11891	HMG box family protein	2.549922783
TTF70519	Rho family GTpase	2.501171814
TTF31287	hypothetical protein TRFO_32279	2.276728765
TTF84547	Dynamin central region family protein	2.237542395
TTF16414	NADP-dependent alcohol dehydrogenase	2.229437347
TTF76816	T-complex protein 1 subunit gamma	2.224989247
TTF58511	lysine–tRNA ligase isoform X2	2.211240873
TTF78798	diphosphate–fructose-6-phosphate 1-phosphotransferase	2.166745694

TTF65307	Serine threonine-protein phosphatase PP-X isozyme 2	2.122385234
TTF71959	Glucose-6-phosphate isomerase	2.114417852
TTF70031	Ras-related protein Rab-15	2.088045028
TTF33037	tRNA binding domain containing protein	2.045065078

Table B3: Fold change of biotinylated *T. foetus* low temperature induced pseudocysts relative to biotinylated trophozoites. All elutions and all experimental trials for each life stage were combined then compared. The samples had all been run on the Orbitrap mass spectrometer using label free and the results were analysed in Peaks [373]. All differentially abundant proteins had a fold change >2 and $p< 0.05$.

Protein ID	Product	Abundance
TTF81436	thioredoxin peroxidase	414868400
TTF65227	adhesin AP65-1 precursor	136023209
TTF51797	cytosolic malate dehydrogenase 1	93423020
TTF11042	actin	82719800
TTF14082	Elongation factor 2	71402020
TTF07843	beta tubulin	69469862
TTF25397	succinate-CoA ligase	44354215
TTF12365	enolase family protein	42847814
TTF44549	MBL fold metallo-hydrolase	38260418
TTF06996	pyruvate:ferredoxin (flavodoxin) oxidoreductase	36933454
TTF81425	Phosphoglycerate kinase	36657695
TTF22965	Malate dehydrogenase	29752200
TTF01814	glyceraldehyde-3-phosphate dehydrogenase	26562274
TTF11905	longation factor 1 alpha	24162330
TTF40594	hydrogenosomal membrane protein 31 precursor	23964945
TTF35950	MBL fold metallo-hydrolase	23064900
TTF35490	heat shock protein 90	20919399
TTF78902	cytoplasmic heat shock protein 70	16924214
TTF26427	pyruvate ferredoxin flavodoxin oxidoreductase	16796331
TTF23112	phosphoglyceromutase	15772091
TTF16414	NADP-dependent alcohol dehydrogenase	15430510
TTF79469	phosphoenolpyruvate carboxykinase	14829839
TTF39915	V-type proton ATPase catalytic subunit A	12537500
TTF18666	actin-like protein	9431274
TTF69995	butanol dehydrogenase	9192480
TTF02727	fructose-1 6-bisphosphate aldolase class II	8456640
TTF16721	40S ribosomal protein S5	8181100
TTF74904	cofilin tropomyosin-type actin-binding protein	8047959
TTF08500	Acetyl-CoA hydrolase	7302390
TTF23613	luminal-binding protein	7104653
TTF34557	thioredoxin-disulfide reductase	7074890
TTF74567	Phosphoglucomutase phosphomannomutase α-β-α domain I family protein	6766330

TTF15197	cytoplasmic heat shock protein 70	6663890
TTF79473	indole-3-pyruvate decarboxylase	6431555
TTF50464	Triosephosphate isomerase cytosolic	6277090
TTF08733	Glucosamine-6-phosphate deaminase	6174930
TTF80890	alpha tubulin	6050140
TTF26092	pyruvate:ferredoxin (flavodoxin) oxidoreductase	5986110
TTF01457	Rpl16ap	5772530
TTF79814	ribosomal protein S3	5663865
TTF55806	60S ribosomal protein L11	5519965
TTF16219	ribosomal protein	5496910
TTF32204	succinyl-CoA ligase [GDP-forming] subunit alpha mitochondrial	5439207
TTF02180	40S ribosomal protein S8	5388335
TTF78798	diphosphate–fructose-6-phosphate 1-phosphotransferase	5342870
TTF04501	actin bundling protein	5313160
TTF50642	40S ribosomal protein S4 X isoform	5273002
TTF08633	adenylyl cyclase-associated protein 1-like	5226086
TTF53150	glucose-6-phosphate 1-dehydrogenase family protein	5194351
TTF36328	Actin-like protein 3	5175300

Table B4: The top 50 most highly abundant proteins identified using TMT mass spectrometry of biotinylated *T. foetus* trophozoites and pseudocysts. Pseudocysts were induced by the lowering of the environmental temperature to 4°C. All elutions, all experimental trials and life stages were combined to give a total peak area (abundance) for all identified proteins.The samples had all been run on the Orbitrap mass spectrometer using label free m/s and the results were analysed in Peaks [373]

.3 Chapter 4 Appendix

Sample	Experiment	No. of Reads	Alignment Rate (%)	Preparation Method
1	Trophozoite Trypsin Wash (30 min) 1	4146360	87.13	Poly-A Selection
2	Trophozoite Trypsin Wash (30 min) 2	7349902	90.64	Poly-A Selection
3	Trophozoite Trypsin Wash (30 min) 3	4788146	93.47	Poly-A Selection
4	Pseudocyst Trypsin Wash (30 min) 1	23525942	0.53	Zymo-kit
5	Pseudocyst Trypsin Wash (30 min) 2	23885184	1.64	Zymo-kit
6	Pseudocyst Trypsin Wash (30 min) 3	28682918	3.36	Zymo-kit
7	Trophozoite Trypsin Wash (1 hour) 1	30250978	95.90	Poly-A Selection
8	Trophozoite Trypsin Wash (1 hour) 2	33638458	95.42	Poly-A Selection
9	Trophozoite Trypsin Wash (1 hour) 3	38894232	8.03	Zymo-kit
10	Pseudocyst Trypsin Wash (1 hour) 1	35507766	7.59	Zymo-kit
11	Pseudocyst Trypsin Wash (1 hour) 2	33849182	1.26	Zymo-kit
12	Pseudocyst Trypsin Wash (1 hour) 3	28381338	95.83	Poly-A Selection
13	Trophozoite Trypsin Wash (24 hours) 1	14634181	8.99	Zymo-kit
14	Trophozoite Trypsin Wash (24 hours) 2	22552918	9.04	Zymo-kit
15	Trophozoite Trypsin Wash (24 hours) 3	51201848	10.29	Zymo-kit
16	Pseudocyst Trypsin Wash (24 hours) 1	30086208	7.13	Zymo-kit
17	Pseudocyst Trypsin Wash (24 hours) 2	37850319	8.36	Zymo-kit
18	Pseudocyst Trypsin Wash (24 hours) 3	32535957	7.15	Zymo-kit
19	Trophozoite PBS Wash (1 hour)	23207498	39.16	Zymo-kit
20	Trophozoite PBS Wash (24 hours)	25562814	92.90	Poly-A Selection
21	Pseudocyst PBS Wash (24 hours	22905766	85.31	Poly-A Selection
22	MDCK cells	36795633	0.04	Zymo-kit
23	DK2 Control 1	21229824	58.08	Zymo-kit
24	DK2 Control 2	51657211	61.60	Zymo-kit
25	DK2 Control 3	31988385	95.66	Poly-A Selection
26	DK2 Supernatant 1	17683526	95.45	Poly-A Selection
27	DK2 Supernatant 2	25688412	95.30	Poly-A Selection
28	DK2 Supernatant 3	26781840	95.16	Poly-A Selection

Table C1: The number of reads produced from RNA-Seq data from *T. foetus* DK2 cells in the presence of an MDCK monolayer and Mapping percentage of the reads against the reference *T. foetus* genome. Samples included several wash steps and controls of cells only of the DK2 strain of *T. foetus*. All RNA samples were produced using the Qiagen RNeasy kit and were sequenced on the Illumina Novaseq. The library preparation was performed either using poly-A selection or a 'zymo' kit. Mapping of the reads to the assembled and annotated *T. foetus* genome was performed using Hisat2 on the Galaxy Server.

Trypsin Upregulated Using DK2 Controls		
Gene ID	Fold change	Gene Product
TTF10220	14.46	Flavodoxin-like fold family protein
TTF83273	14.17	Hypothetical
TTF84480	13.21	β-galactosidase
TTF60835	12.79	CAMK protein kinase
TTF43440	12.72	Iron only hydrogenase large subunit C-terminal containing protein
TTF31570	12.63	Hypothetical
TTF81526	12.54	6-phosphoglucolactonase
TTF41322	12.27	ABC-transporter protein
TTF81248	12.08	CAMK protein kinase
TTF83248	11.96	CAMK protein kinasel
Trypsin Downregulated Using DK2 Controls		
Gene ID	Fold change	Gene Product
TTF81579	-12.79	DnaJ containg protein
TTF45476	-11.96	Potassium-sodium hyperpolarisation activated gated channel
TTF34734	-11.91	Cysteine protease
TTF34777	-11.29	Cysteine protease
TTF64110	-11.25	Cysteine protease
TTF03418	-10.72	40S ribosomal protein S11
TTF74559	-10.45	Sialidase-like protein
TTF51463	-10.42	Myb-like DNA-binding domain containing protein
TTF70500	-9.89	Serine-threonine protein kinase 2
TTF48862	-9.89	Hypothetical

Table C2: Differential expression of *T. foetus* DK2 cells after their addition to an MDCK monolayer. The cells were removed from the monolayer using a trypsin wash and RNA extracted. The top 10 preferentially up and down regulated genes are relative to *T. foetus* DK2 controls are shown.

PBS Upregulated with DK2 Controls		
Gene ID	Fold change	Gene Product
TTF83273	15.11	Hypothetical
TTF25670	14.58	Dihydropyrimidase
TTF84480	13.85	β-galactosidase
TTF43440	13.28	Iron-only hydrogenase large subunit C terminal domain containing protein
TTF60835	13.18	CAMK family protein
TTF80312	13.09	Hypothetical
TTF69994	13.07	Hypothetical
TTF81526	12.59	6-phosphogluconolactase
TTF81248	12.56	CAMK family protein
TTF23280	12.45	NA
PBS Downregulated with DK2 Controls		
Gene ID	Fold change	Gene Product
TTF45476	-14.09	Potassium-sodium hyperpolarisation activated gated channel
TTF60658	-12.67	Pyrroline-5-carboxylate reductase
TTF48862	-11.94	Hypothetical
TTF42967	-11.50	Acetlyhydrolase
TTF74496	-10.83	Zinc finger protein
TTF42134	-10.78	SEl 1 protein
TTF42966	-10.69	Acetlyhydrolase
TTF46664	-9.75	Poxvirus D5 protein
TTF13115	-9.69	Clan CA family cysteine peptidase
TTF23726	9.37	NA

Table C3: Differential expression of *T. foetus* DK2 cells after their addition to an MDCK monolayer. The cells were removed from the monolayer using a PBS wash prior to the trypsin was and RNA extracted. The top 10 preferentially up and down regulated genes, relative to *T. foetus* DK2 controls are shown.

Supernatant Upregulated Using DK2 Controls		
Gene ID	Fold change	Gene Product
TTF60835	15.73	CAMK family protein kinase
TTF80312	15.32	Hypothetical
TTF84480	14.93	β-galactosidase
TTF43440	14.03	Iron only hydrogenase large subunit C-terminal containing protein
TTF19166	13.60	Bowman birk serine protease inhibitor family protein
TTF25670	13.59	Dihydropyrimidinase
TTF83273	13.50	Hypothetical
TTF40347	13.37	Small-GTP binding protein
TTF31013	12.73	Clan SB family serine peptidase
TTF83251	12.72	Myb-like DNA-binding domain containing protein
Supernatant Downregulated Using DK2 Controls		
Gene ID	Fold change	Gene Product
TTF45476	-11.04	Potassium-sodium hyperpolarisation activated gated channel
TTF48862	-10.42	Hypothetical
TTF75464	-10.16	Cysteine protease
TTF64908	-9.89	Ser-Thr protein phosphatase
TTF41202	-9.57	Hypothetical
TTF44561	-9.49	Dnak protein
TTF31045	-9.46	Hypothetical
TTF07998	-9.24	Hypothetical
TTF36147	-9.21	Sugar O-acetyltransferase
TTF74467	-9.16	Hypothetical

Table C4: Differential expression of *T. foetus* DK2 cells after their addition to an MDCK monolayer. The supernatant from the experiment was removed from the monolayer and RNA extracted. The top 10 preferentially up and down regulated genes relative to *T. foetus* DK2 controls are shown.

.4 Chapter 5 Appendix

Gene	All Trichomonads			*T. foetus* strains only		
	Non-synonymous	Synonymous	Total	Non-synonymous	Synonymous	Total
TTF00576	0	3	3	0	0	0
TTF00910	2	4	6	0	0	0
TTF01152	3	0	3	0	0	0
TTF01161	2	3	5	0	0	0
TTF02039	9	9	18	0	0	0
TTF02745	12	4	16	2	0	2
TTF03161	9	4	13	0	0	0
TTF03169	10	7	17	2	0	2
TTF04635	56	21	77	27	8	35
TTF05043	5	8	13	0	0	0
TTF05727	2	4	6	0	0	0
TTF05864	18	5	23	2	0	2
TTF08006	7	7	14	1	0	1
TTF08916	0	4	4	0	0	0
TTF09063	8	22	30	1	0	1
TTF09721	3	6	9	0	2	2
TTF11197	79	42	121	2	1	3
TTF12012	11	5	16	0	0	0
TTF13670	11	14	25	0	0	0
TTF13877	26	22	48	1	0	1
TTF14449	1	1	2	1	0	1
TTF14901	22	31	53	1	4	5
TTF16365	18	21	39	4	0	4
TTF16447	8	3	11	0	0	0
TTF18002	14	12	26	0	1	1
TTF19171	12	12	24	10	8	18
TTF20128	7	4	11	0	0	0
TTF21569	5	6	11	0	1	1
TTF21670	7	5	12	1	0	1
TTF23071	32	20	52	6	0	6

TTF25642	7	9	16	2	0	2
TTF28036	22	21	43	1	0	1
TTF28414	8	4	12	0	0	0
TTF29007	5	9	14	0	1	1
TTF31219	4	3	7	2	2	4
TTF33095	4	4	8	0	0	0
TTF33122	11	3	14	0	0	0
TTF33823	3	3	6	1	0	1
TTF33905	3	3	6	1	0	1
TTF34783	14	8	22	5	0	5
TTF37306	11	2	13	2	0	2
TTF38179	16	23	39	3	2	5
TTF40724	15	20	35	1	0	1
TTF43413	12	28	40	3	0	3
TTF43627	12	7	19	1	0	1
TTF44528	43	25	68	2	0	2
TTF44600	10	17	27	5	1	6
TTF44685	11	8	19	2	0	2
TTF44949	6	11	17	2	0	2
TTF45042	201	64	265	5	3	8
TTF45771	7	4	11	2	0	2
TTF46221	39	19	58	1	0	1
TTF47476	5	9	14	0	0	0
TTF48521	4	5	9	0	0	0
TTF49933	16	11	27	2	0	2
TTF50636	6	10	16	0	0	0
TTF53402	6	21	27	0	0	0
TTF56587	6	8	14	3	0	3
TTF62174	33	11	44	1	0	1
TTF62376	1	3	4	0	0	0
TTF63395	27	8	35	3	0	3
TTF63680	6	15	21	1	1	2
TTF67401	2	4	6	0	1	1
TTF70906	6	13	19	1	0	1

TTF70954	4	8	12	2	1	3
TTF71197	34	52	86	5	0	5
TTF72511	5	2	7	0	0	0
TTF72966	26	20	46	5	0	5
TTF73221	27	25	52	4	2	6
TTF73824	11	27	38	0	1	1
TTF74586	17	7	24	0	0	0
TTF74797	2	2	4	0	0	0
TTF75601	48	43	91	8	1	9
TTF79943	14	18	32	1	0	1
TTF80312	2	1	3	2	1	3
TTF82375	2	1	3	1	0	1
TTF83271	1	12	13	0	1	1

Table D1: Number of SNPs identified in shortlist produced in Chapter 4. The numbers were identified when five *T. foetus* and one *T. mobilensis* strain were compared to the reference *T. foetus* genome ('All trichomonads') and when only the five *T. foetus* strains were compared to the reference ('*T. foetus* only'). For each gene the number of non-synonymous and synonymous mutations are given.